1-1-2002

Small sample item parameter estimation in the three parameter logistic model : using collateral information.

Lisa A. Keller
University of Massachusetts Amherst

SMALL SAMPLE ITEM PARAMETER ESTIMATION IN THE THREE PARAMETER LOGISTIC MODEL: USING COLLATERAL INFORMATION

A Dissertation Presented

By

Lisa A. Keller

Submitted to the Graduate School of the
University of Massachusetts Amherst in partial fulfillment
of the requirements for the degree of

DOCTOR OF EDUCATION

September 2002

School of Education

SMALL SAMPLE ITEM PARAMETER ESTIMATION IN THE THREE PARAMETER LOGISTIC MODEL: USING COLLATERAL INFORMATION

A Dissertation Presented

By

Lisa A. Keller

Approved as to style and content by:

Hariharan Swaminathan, Chair

Ronald K. Hambleton, Member

Arnold D. Well, Member

Andrew Effrat, Dean
School of Education

DEDICATION

To Fraser, my patient and loving husband

ACKNOWLEDGEMENTS

There are so many people that have made the completion of this work possible, and their contributions have been invaluable. There is no way to adequately thank them for all the support, guidance, wisdom and insight that they have provided. I would like to start by thanking my committee for their support. I am especially grateful to my advisor, Professor Swaminathan, whose calm, collected and rational demeanor allowed me to struggle through the rough times and to not give up. His patience and encouragement were immense; I am forever appreciative. Additionally, the comments and insight provided by Professor Ron Hambleton have helped shaped this work into a cohesive document.

In addition to helping me produce this work, Swami and Ron have been instrumental in my development as a researcher. Their patience and interest in my endeavors have provided me with many opportunities to grow, both personally and professionally. I cannot fail to thank Professor Stephen G. Sireci, who encouraged me to study psychometrics, and who has always been there to give advice and create opportunities. I have been lucky to have such a faculty to work with. I am indebted to all three of my professors.

In addition to the faculty support, the friendship and loyalty of my fellow graduate students has gotten me through some of the darkest times. I thank Michael Jodoin for his never-ending willingness to listen to my ramblings about parameter estimation, his readiness to offer insight, and most of all his dedication to maintaining the caffeine-level in my bloodstream. He has offered friendship and kindness in addition to

his professional support. April Zenisky has been with me since day one and her hugs, smiles, cards and outings have provided me with the emotional support essential to surviving graduate school. In addition, Mary Pitoniak and Billy Skorupski made me laugh, listened patiently and provided friendship throughout.

I also would like to thank Educational Testing Service for providing the funding with which to complete this study. Additionally, I would like to thank Tim Davey for his insight and thoughtful contributions to the project.

Lastly and most importantly, I am grateful to the patience and encouragement of my family. All the sacrifices they have made in the name of this document, and my education as a whole cannot fit on one page. I thank them for understanding and letting me do what needed to be done. I especially thank Fraser, who believed in me when I could not believe in myself. He provided the strength for both of us to make it through, and a home where I could always go where being myself was always enough. For the countless hours of listening to me and consoling me when things looked bleakest, I thank him with all my heart.

ABSTRACT

SMALL SAMPLE ITEM PARAMETER ESTIMATION IN THE THREE PARAMETER LOGISTIC MODEL: USING COLLATERAL INFORMATION

SEPTEMBER 2002

LISA A. KELLER, B.S., ST. MICHAEL'S COLLEGE

M.S. UNIVERSITY OF MASSACHUSETTS AMHERST

Ed. D. UNIVERSITY OF MASSACHUSETTS AMHERST

Directed by Professor Hariharan Swaminathan

The appeal of computer adaptive testing (CAT) is growing in the licensure, credentialing, and educational fields. A major promise of CAT is the more efficient measurement of an examinee's ability. However, for CAT to be successful, a large calibrated item bank is essential. As item selection depends on the proper calibration of items, and accurate estimation of the item information functions, obtaining accurate and stable estimates of item parameters is paramount. However, concerns of item exposure and test security require item parameter estimation with much smaller samples than is recommended. Therefore, the development of methods for small sample estimation is essential.

The purpose of this study was to investigate a technique to improve small sample estimation of item parameters, as well as recovery of item information functions by using auxiliary information about item in the estimation process. A simulation study was conducted to examine the improvements in both item parameter and item information

recovery. Several different conditions were simulated, including sample size, test length, and quality of collateral information. The collateral information was used to set prior distributions on the item parameters. Several prior distributions were placed on both the *a*- and *b*- parameters and were compared to each other as well as to the default options in BILOG.

The results indicate that with some relatively good collateral information, nontrivial gains in both item parameter and item information recovery can be made. The current literature in automatic item generation indicates that such information is available for the prediction of item difficulty. The largest improvements were made in the bias of both the *a*-parameters and the information functions. The implications are that more accurate item selection can occur, leading to more accurate estimates of examinee ability.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

INTRODUCTION

Item response theory (IRT) serves as the cornerstone of modern educational testing technology. The advantages of using item response theory in testing are well documented (Hambleton & Swaminathan, 1985). Among the major advantages of item response theory over classical test theory based procedures is that item parameters can be obtained independently of the examinee population that takes the test and the abilities of examinees can be determined independently of the set of items taken and compared. This second feature makes computer adaptive testing (CAT) possible. In a CAT framework, an examinee is administered an item that provides the most information at the examinee's ability level; testing continues until the ability of an examinee is determined to the desired degree of precision. Unlike in conventional testing, in a CAT, different examinees are administered different sets of items, and since the abilities of the examinees are on a common scale, the examinees can be compared. This design results in very efficient test administration and is currently employed in several large scale testing programs.

Among the promises of CAT is more efficient estimation of the candidate's ability. However, before CAT can be implemented, the item parameters need to be estimated. During the CAT administration, these item parameter estimates are treated as true values, and the ability of an examinee is estimated.

The rest of this chapter proceeds by detailing the role of item parameters in CAT. A statement of the problem and the purpose of the current study follow this discussion.

1.1 Role of Item Parameters in CAT

In order for the item parameters to be estimated adequately, large samples of examinees are necessary, especially as model complexity increases. Hambleton and Swaminathan (1985) recommend 1000 examinees for the three-parameter model, which is the most complex of the dichotomous item response models, and will be described in detail in the next chapter. In a CAT environment, concerns for test security, and hence item exposure, make it difficult, if not impossible, to obtain such samples. However, as items are chosen for administration based on the item parameters, proper estimation of these parameters is essential. Furthermore, Hambleton and Jones (1994) showed that the use of imprecise item parameters lead to an overestimate of test information, yielding ability estimates that are less accurate than they appear. Given the importance of the item parameters, it is necessary to develop methods to obtain acceptable item parameter estimates with small samples.

Item parameters are used not only to estimate the ability of an examinee, but also to select the items that are used for that purpose. Most item selection algorithms rely on the item information as a basis for selection. The item information is computed using all item parameters, of course, however, the *a*-parameter plays an important role in the calculation of the item information. Indeed, the amount of information contained in an item is proportional to the square of the *a*-parameter. Therefore, while the estimation of all item parameters is important the proper estimation of the *a*-parameter is crucial to item selection and the proper estimation of item information.

Furthermore, the standard error of the resulting ability estimate is the inverse of the test information function. Thus, to adequately determine the proper standard error of the ability estimate, the information function must be adequately estimated. In many cases, the *a*-parameter is overestimated, especially in small samples, which would result in an overestimation of the item information function. As mentioned above, the items with the highest *a*-parameters are often chosen for administration, resulting in the choice of items whose *a*-parameters may be overestimated. Since the standard error of the ability estimate is based on the *test* information, the accumulation of this error across several items may lead to a gross overestimation of the test information function, and hence a substantial underestimate of the resulting standard error, leading to the conclusion that the ability estimate is adequately precise. The worse the estimation of the *a*-parameter, the more gross the error in information. Therefore, the benefit of more efficient estimation of ability may not be realized when the item parameters are poorly calibrated. Given the importance of a-parameter in this role, the proper estimation of this parameter, and most importantly, the resulting information function, is of central concern.

Despite the importance of proper estimation of the *a*-parameter, the *b*-parameter also places and important role in the item selection, and as such, must be properly estimated as well. However, the *b*-parameter is the most easily estimated parameter, and as such the situation is less critical. Nonetheless, the *b*-parameter is used to match the difficulty of the item with the ability estimate of the candidate, which allows for the efficiency of CAT. Therefore, the estimation of the *b*-parameter is also important for item selection. In the event that the *b*-parameter is

underestimated, the item is taken to be easier than it actually is, and when administered, may be too difficult for the examinee. While this may be of psychometric concern, it also is of psychological concern; administering items that are too difficult for an examinee can lead to increased anxiety, resulting in poorer performance than is warranted by the candidate's ability. However, the gravity of this situation is less serious, as it is unlikely that any one candidate is given a series of items whose *b*-values are underestimated.

Perhaps most importantly, the efficiency of the estimation promised by CAT can only exist when the item parameters are properly calibrated. As the efficiency in estimation relies on the matching of items and ability, if item parameters in general are not well estimated, the resulting provisional estimates of ability are also not very accurate, resulting in a loss of efficiency in the testing procedure. Furthermore, while the bias of the *b*-parameters may be of less concern than the bias of the *a*-parameters, it is still a matter for concern if the bias of the parameters is large. Again, to the extent that an adequate match is not made between the candidate's ability and the difficulty of the item, the efficiency of the estimation is not realized.

1.2 Statement of Problem

While there has been some research on small sample estimation, with different methods yielding minor improvements in estimation, the need for better methods exists. The literature offers very few alternatives for practitioners. The bulk of the research has focused on modifying existing item response models to limit the demands placed on estimation, or obtaining optimal samples for calibration. However, these

alternatives do not provide for methods that are ideal. In the case of modified models, the estimation is limited by fixing one or more item parameter. While this may limit the demands of calibration, the resulting models do not retain the flexibility of the original model to adequately reflect the data. In terms of optimal sampling methods, most of the proposed methods require knowing the true item parameters and/or the ability parameters. The one exception is Slater (2001), which is discussed in more detail in chapter 2. The results of these methods are not only impractical, but often provide little or no improvement in estimation. Therefore, alternative methods for reducing the necessary sample sizes are needed. Some promise is shown in the use of collateral information in estimation, and as such is a line of study worth pursuing.

Swaminathan, Hambleton, Sireci, Xing and Rizavi (in press) and Mislevy (1986) considered using additional information about items to aid in the estimation process. Mislevy (1986) discussed using item features to aid in the estimation process (e.g. number of words, item format, cognitive processes), while Swaminathan et al. (in press) considered using expert judgments about the difficulty of items and incorporating this information, via item-specific priors, in the estimation of item parameters. These approaches have shown some success, especially in the estimation of the a- and c-parameters, which are typically more difficult to estimate. As mentioned above, improving the estimation of the *a*-parameter is of central concern, and thus this method is very promising. Additionally, by recovering both the *a*- and *c*-parameters more successfully, undoubtedly the item information functions would have been better estimated as well. Although the study did not consider the improvement in estimation of the information function, by improving the estimation of all parameters,

and most importantly the *a*-parameter, the combined improvement would surely lead to a much more accurate estimate of information, leading to a more accurate ability estimate. The one limitation of the Swaminathan et al. study is in the costly attainment of expert judgments. If the same results could be attained using a more readily source of item information, this method would be a very practical approach to improved estimation.

Work in automated item generation (AIG) has also lead to some promising approaches to estimating item parameters. Several studies (Embretson, in press; Enright et al., 1999; Dennis et al., in press) have investigated the feasibility of *predicting* item parameters from various item features (e.g. number of words, cognitive processes), to reduce the need for item pre-calibration. Since the demand for a large number of items requires the production of items with known parameters, this growing line of research seeks methods to produce items with parameters which can be predicted accurately enough so that the predicted parameters can be used as the item parameters, and pre-calibration can be eliminated. While this seems optimistic, it is promising, and the methodology developed in this area can be used in conjunction with other methods to at least reduce the sample sizes necessary to accurately estimate item parameters, if not eliminate the need for calibration.

By combining the work of Mislevy, Swaminathan et al. and the work in AIG, a promising approach for small sample calibration emerges. Using predicted item parameters as additional information about items has the potential for improving the accuracy of item parameter estimation in small samples.

1.3 Purpose of Study

The purpose of the study is to conduct a simulation study to investigate methods for small sample estimation that can be implemented in an operational CAT setting. First, a literature review to examine the potential of the different estimation techniques was conducted, as well as to examine the previous attempts at small sample estimation.

Since the proposed method will rely on the prediction of item parameters, an investigation into the feasibility of predicting item parameters, as well as what type of information is available to predict item parameters was conducted. Once it has been determined which item parameters can be successfully predicted from which item features, this information can be incorporated into the estimation process through the use of Bayesian estimation techniques.

The proposed method includes obtaining information on items to aid in the estimation process. Since pretest items (whose parameters are unknown) are typically administered simultaneously with operational items (whose parameters are known), this auxiliary information about the items can be used to predict the item parameters, which can aid in the estimation of the item parameters by allowing for the specification of prior distributions for the appropriate item parameters for each item. These prior distributions will aid in the estimation of the item parameters by restricting the range of possible estimates to those that are most likely.

The document will proceed by providing a review of the major item estimation techniques, followed by a description of previous attempts at small sample estimation. Efforts at predicting item parameters, particularly in the framework of automated item

generation are reviewed along with a description of the types of collateral information that can be obtained. Lastly, previous attempts to incorporate this information in the estimation process are detailed.

Chapter 2

LITERATURE REVIEW

This chapter begins with a description of the item response model that is used in this study. It continues by describing methods for item parameter estimation, and attention is drawn to their feasibility in the realm of small sample estimation. Next, previous attempts in small sample estimation are presented along with the relative success of each of the proposed methods. Following that, a look into the automatic item generation literature provides some information regarding the feasibility in predicting item parameters, as well as the type of information that can be used to do so. The chapter ends with a summary of the reviewed literature and an explanation of how the work in AIG can be combined with item estimation methods to potentially reduce the required samples for accurate estimation of item parameters as well as item information functions.

2.1 The Item Response Model

Item response theory postulates a probabilistic relationship between an examinee's unobserved ability θ, the characteristics of an item (item parameters) and the observed dichotomous response (U) to the item. While the probability of a correct response , U=1, can be modeled through any probability distribution function, the most commonly used function is the logistic function. The number of parameters that characterize an item characterizes the resulting item response model; in the one-parameter (1PL) or the Rasch model, the item is characterized by one parameter, the

item difficulty, b_j. The two-parameter model (2PL) is characterized by the item difficulty b_j and the item discrimination parameter, a_j. In the three-parameter model (3PL), a lower asymptote, c_j , is introduced to take into account "guessing" on the item. The most general three-parameter item response model is given as

$$P(u_{ij}=1)=c_j+(1-c_j)\frac{e^{1.7a_j(\theta-b_j)}}{1+e^{1.7a_j(\theta-b_j)}}.$$

The two-parameter model is obtained by setting $c_j = 0$, while the one-parameter model is obtained by setting $c_j = 0$ and a_j=1.

Since the goal of any testing program is to provide estimates of an examinee's ability on a given trait, the selection of the correct item response model is a critical step in computing the appropriate estimate of ability. It is necessary to choose the model which best describes the data. Ideally, several IRT models should be fit to the data, and the model exhibiting best fit should be selected. However, in the case of multiple-choice data, empirical studies have shown that the 3PL best models the data. That is not surprising, given that there is a chance that examinees guess on items, leading to a need for a lower asymptote. As the majority of tests are largely composed of multiple-choice items, the 3PL will serve as the focus of this paper.

Once the item response model is selected, the item parameters must be estimated. There are several estimation procedures used to estimate the item parameters, when both item and ability parameters are unknown. Among the most popular methods are joint maximum likelihood (JML) where the item and ability parameters are estimated jointly, marginal maximum likelihood (MML) procedures

where the ability distribution is integrated out, and Bayesian procedures. The most commonly implemented Bayesian procedure is Bayes Modal Estimation (BME). Each of these methods is described below.

2.2 Non-Bayesian Estimation Procedures

In non-Bayesian estimation procedures the parameter estimates are obtained based solely on the information contained in the response patterns of the examinees. Therefore, these procedures are completely objective. It is for this reason that some practitioners prefer non-Bayesian techniques, as Bayesian procedures require assumptions about the distributions of the parameters. Of the non-Bayesian techniques, JML and MML are the most popular, and will be described briefly below.

2.2.1 Joint Maximum Likelihood

As the name implies, in joint maximum likelihood, the item and ability parameters are estimated simultaneously. Maximum likelihood procedures are used on the joint likelihood to find the maximum likelihood estimates of both item and ability parameters. In the case of dichotomous models, which are of interest in this study, given a response vector for a person $u= (u_1, u_2, \ldots, u_n)$ to n dichotomous items, the likelihood function for N examinees responding to n items is expressed as:

$$L(u_1,u_2,\ldots,u_N \mid \theta,a,b,c) = \prod_{i=1}^{N}\prod_{j=1}^{n} P_{ij}^{u_{ij}} Q_{ij}^{1-u_{ij}}$$

where u_i is the response *vector* for examinee i, P_{ij} is the probability of a correct response of person i to item j, as given by the 3P model, and $Q_{ij}=1-P_{ij}$.

Clearly, as the number of items and the number of examinees increase, so does the complexity of the likelihood function. For each item (in the case of the 3PL), there are 3 parameters that need to be estimated, and for each examinee there is one ability parameter. Therefore, there are 3n+N parameters that need to be estimated.

As indicated by Hambleton, Swaminathan, and Rogers (1991), before these estimates can be determined, there is the problem of scale indeterminacy that requires resolution. In the 3PL, for instance, if θ is replaced by $\theta^* = P(\alpha\theta+\beta)$ then $P(\theta) = P(\theta^*$). Given that α and β are arbitrary scaling constants, there is no unique maximum. Therefore, in order to obtain unique solutions, constraints need to be imposed on the equations. This situation is commonly remedied by fixing the mean and standard deviation of the ability parameters to zero and one, respectively. Once this scale is set, then the estimates can be determined.

Since both item and ability parameters are estimated simultaneously, a "divide-and-conquer" strategy is imposed. Initial estimates are placed on the theta values. Typically a logit percent correct score is used. These theta values are treated as known, and the item parameters are estimated by finding the set of item parameters that maximize the likelihood surface, given the theta values used. The maximization is accomplished by taking the first derivative of the likelihood function with respect to each parameter, setting it equal to zero, and solving. Since these equations are often impossible to solve in closed-form, numerical techniques, such as Newton-Raphson, are employed to obtain the maximum. After the item parameters are estimated, they are treated as known, and the ability parameters are then similarly estimated. Once the ability parameters are re-estimated, they are fixed and item parameters are re-

estimated. This process is continued until there is little difference between stages of estimation.

While there does seem to be elegance in this method due to the lack of distributional assumptions, there are some fairly major criticisms. Estimates for examinees with perfect scores, or zero scores are impossible to obtain. Similarly, estimating parameters for items which all examinees get right/wrong is impossible. Therefore, the removal of these cases is necessary in order to proceed (Hambleton, Swaminathan &Rogers, 1991).

Since the item parameters and the ability parameters are estimated simultaneously, the JML estimates in the 3PL case are not consistent. Neyman and Scott (1948) showed that large numbers of incidental parameters could affect the consistency of the estimates of structural parameters. In the case of item parameter estimation, the ability parameters are considered incidental, or nuisance, parameters, and the item parameters are structural parameters. As the number of examinees increases, the number of incidental parameters increases, and the consistency of the estimate becomes suspect. However, Swaminathan and Gifford (1983) showed that consistent estimates of item parameters in the 3PL are possible if both the number of items and examinees becomes large. In the case of CAT, this is not feasible. Additionally, unless restrictions are placed on the values that the item parameters can take, numerical procedures will often fail (Hambleton, Swaminathan & Rogers, 1991).

The problem of improper estimates can be remedied by placing the necessary restrictions on the values that the item and ability parameters can take. Swaminathan and Gifford (1982, 1985, 1986) developed a series of Bayesian procedures that set

prior distributions on the parameters resulting in proper estimates. However, while these priors aided in the proper estimation, it did not aid in obtaining consistent estimates when large samples and long tests are not available.

To remedy the problem of inconsistency, it is necessary to estimate the item parameters independently of the ability parameters. Integrating out the ability parameter, or, marginalizing the joint distribution can accomplish this. The next section talks about The marginal maximum likelihood procedure, which does precisely that, is described below.

2.2.2 Marginal Maximum Likelihood Procedure

Marginal maximum likelihood estimation also allows for the estimation of the item parameters when both the item and ability parameters are unknown. Given the joint density of the parameters (both item and ability), the marginal density of the item parameters can be obtained by integrating out the ability parameter, θ. This marginal density gives rise to the marginal likelihood function. This function can then be maximized and item parameter(s) are given as the solution(s) to the likelihood equations. Once these item parameters are obtained, they are taken to be the true item parameters, and the ability parameter is then estimated, using any of the ability estimation techniques available. These estimates have been shown to be more accurate than those obtained using JML (Seong, 1990).

More specifically, the probability of an examinee j obtaining a particular

response pattern U is given by:

$$P[U \mid \theta, a, b, c] = \prod_{i=1}^{n} P_i^{U_i} (1 - P_i)^{1-U_i}$$

where P_i is given by the item response function above, and n is the number of items administered. It follows that :

$$P[U, \theta \mid a, b, c] = \prod_{i=1}^{n} P_i^{U_i} (1 - P_i)^{1-U_i} g(\theta)$$

(Hambleton & Swaminathan, p. 140). Hence

$$\pi_u = P[U \mid a, b, c] = \int_{-\infty}^{\infty} \prod_{i=1}^{n} P_i^{U_i} (1 - P_i)^{1-U_i} g(\theta)\, d\theta$$

(Hambleton & Swaminathan, p.140).

This quantity, π_u, is the marginal probability of obtaining response pattern u. Note that there are 2^n possible response patterns for the n items, therefore if there are r_u examinees that obtain response pattern u, then the likelihood function is given by:

$$L \propto \prod_{u=1}^{2^n} \pi_u^{r_u}$$

and taking the logarithm:

$$\ln L = c + r_u \sum_{u=1}^{2^n} \ln \pi_u$$

where c is a constant. Differentiating and solving the resulting likelihood equations yields the marginal maximum likelihood estimate of the item parameters.

Marginal maximum likelihood estimates have the benefit of being consistent estimators, provided that the item response model and the distribution, $g(\theta)$, is chosen appropriately (Harwell & Baker, 1991). The property of consistency is asymptotic, though, and as such implies that with small samples, this consistency property may not

be achieved. Further, correctly specifying $g(\theta)$ requires that the distribution of the examinee population is known. This is only realistic if large numbers of examinees are used. Therefore, in the case of small sample estimation, MML estimates may not be optimal either.

Depending on the data set being analyzed, the MML estimates of item parameters may assume unreasonable values (Mislevy, 1986). This may be especially true in the case of small samples, where less information is available to estimate item parameters. The use of Bayesian techniques in these instances can limit the possible range of values a parameter can attain through the specification of a prior distribution. Therefore, a consideration of Bayesian estimation techniques is warranted.

2.3 Bayesian Estimation Procedures

Bayesian estimation procedures employ one general principal. Information about the distribution of item parameters is used in the estimation process to obtain more accurate estimates. In many cases, there is enough information about the item to be able to do this in a reasonable manner. As indicated in Lord (1986), one clear advantage of Bayesian methods is that the posterior *mean* minimizes the overall mean squared error (MSE) of estimation, provided that appropriate prior distributions are used. One consequence of this reduced error, however, is the acceptance of increased bias. While this same property is not true for the *mode* of the estimate (unless, of course, the mean and mode are identical), it may be close enough to the mean to be acceptable. Additionally, O'Hagan (1976) showed that the marginal posterior mode is preferred over the joint posterior mode as an approximation to the posterior mean. In

this study, we will consider Bayes Modal estimation, which produces a point-estimate that is the mode of the posterior. A description of the method is provided below.

Regarding small sample calibration, Swaminathan and Gifford (1986) showed that Bayesian procedures produced more accurate estimates than did joint maximum likelihood procedures.

2.3.1 Bayes Modal Estimates

In obtaining Bayes Modal estimates, a process similar to that used in MML estimation is used. In this instance, however, prior distributions are placed on the item and ability parameters. These prior distributions reflect the a priori belief about the distribution of the item parameters. As the amount of information available about the distribution of the parameters differs in each case, the choice of prior distributions can reflect the amount of confidence placed in the information. Since the choice of prior distribution does affect the resulting estimate, a strong prior reflects great confidence in the information being used in the estimation process. Prior distributions can be placed on any or all parameters. Once the prior distributions are determined, the posterior distribution is obtained by multiplying the likelihood by the prior distribution(s), and can be used to make inferences about the desired parameters.

More specifically, the likelihood function is obtained as above,

$$L(\boldsymbol{u}\,|\,\theta,\boldsymbol{a},\boldsymbol{b},\boldsymbol{c}) = \prod_{i=1}^{N}\prod_{j=1}^{n} P_{ij}^{u_{ij}} Q_{ij}^{1-u_{ij}}$$

where

$$\boldsymbol{\theta} = (\theta_1 \;\; \theta_2 \;\ldots\; \theta_N)$$
$$\boldsymbol{a} = (a_1 \;\; a_2 \;\ldots\; a_n)$$
$$\boldsymbol{b} = (b_1 \;\; b_2 \;\ldots\; b_n)$$
$$\boldsymbol{c} = (c_1 \;\; c_2 \;\ldots\; c_n)$$

and $\boldsymbol{u} = (u_{11} \;\; u_{12} \;\ldots\; u_{Nn})$ is the vector of the observed responses of N examinees to n items.

If we consider the joint prior density of the item and ability parameters to be $f(\theta, \boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c})$, then the joint posterior distribution of the parameters, given the observed responses can be expressed, via Bayes' Theorem, as

$$f(\theta, \boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}|\, \boldsymbol{u}) \propto L(\boldsymbol{u}|\, \theta, \boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}) f(\theta, \boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c})$$

(Swaminathan & Gifford, 1986).

Since prior distributions are placed on the item parameters, the prior distributions themselves involve parameters. The parameters that are involved in the prior distributions are referred to as *hyperparameters,* and should also be made explicit. For example, the prior distribution of person's ability, θi, is typically taken to be normally distributed. Given the assumption that all person's abilities are independent and identically distributed, the hyperparameters would consist of the common and mean and variance of the normal distribution from which the abilities are drawn.

That is,

$$\theta_i \sim N(\mu, \sigma^2) \text{ for all } i.$$

Here, μ and σ^2 are the hyperparameters. Given this conceptualization, the joint density of the item parameters can be more appropriately expressed as

$$f(\theta, a, b, c, \tau, \eta) = f(\theta \mid \tau) f(a, b, c \mid \eta) f(\tau) f(\eta)$$

where τ is the vector of hyperparameters for the item prior distributions and η is the vector of hyperparameters for the ability parameters. The resulting posterior distribution can be expressed as

$$f(\theta, \tau, a, b, c, \eta) = L(Y \mid \theta, a, b, c) f(\theta \mid \tau) f(\tau) f(a, b, c \mid \eta) f(\eta)$$

(Harwell & Baker, 1991)

It is common to assume that the item parameters, as well as the ability parameters are independent. Therefore, the prior distribution can be written as follows:

$$f(\theta, a, b, c) = f(\theta)\, f(a)\, f(b)\, f(c).$$

Given the joint posterior of the item and ability parameters, point estimates can be obtained. Bayes modal estimates (BMEs) are obtained by finding the mode of the posterior distribution. Just as in the maximum likelihood case, estimates of the item and ability parameters can be obtained from the joint posterior distribution (e.g. Swaminathan & Gifford, 1982, 1985, 1986) or marginalized estimated of the item parameters can be found by integrating out the ability parameter (Mislevy, 1986; Tsutakawa & Lin, 1986). O'Hagan (1976) provides numerical evidence for the

superiority of marginalized solutions, and hence, in this paper, the marginalized approach will be followed.

Integrating out the ability parameter yields the following marginalized posterior:

$$f(a,b,c,\tau) \propto \int_{\eta}\int_{\theta} L(Y \mid a,b,c,\theta) f(\theta \mid \tau) f(\tau) f(\eta) d\theta d\tau \propto L(Y \mid a,b,c,\tau) f(a,b,c) f(\tau)$$

It is important to note the integrating over the values of theta eliminates the dependence of the posterior on theta, but not the hyperparameters contained in τ. Additionally, integrating over the population distribution of item parameters has not eliminated the need to specify values for the hyperparameters η (Harwell & Baker, 1991).

Using the marginalized posterior distribution, the maximum of this density is found by taking partial derivatives with respect to the item parameters, and setting them equal to zero. As it is typically impossible to solve these equations, numerical procedures are necessary. In this instance, the resulting equations are typically solved one item at a time in the M (maximization) step of the EM (expectation maximization) algorithm (Mislevy, 1986).

In this instance, if the IRT model and the prior distribution of θ is correct, the resulting item parameter estimates are consistent (Harwell & Baker, 1991). In small sample cases, Harwell and Janosky (1991) showed that BMEs showed less estimation error than MMLEs.

2.3.2 Specification of the Prior Distribution(s)

In order to implement the Bayesian procedures described above, it is necessary to specify the prior distributions of the item parameters. In the context of IRT, many authors have suggested informative priors to be set on the item parameters (Mislevy, 1986; Mislevy & Stocking, 1989; Swaminathan & Gifford, 1985, 1986). A prior distribution is considered informative if its variance is small (Harwell & Janosky, 1991), as the implication is that the value of the parameter will be clustered tightly to the specified mean of the prior distribution. The effect of an informative prior is to "shrink" the estimate toward the mean of the prior by an amount proportional to the information contained in the prior distribution (Mislevy & Stocking, 1989). If the variance of the prior distribution is large, the clustering affect is small, and hence has less affect on the parameter estimation. Such a prior is referred to as "non-informative." Therefore, the determination of the variance of the prior distribution plays a major role in the estimation of item parameters, and as such must be chosen carefully. There are two central issues to consider in the determination of the prior distributions: which parameters require priors, and what form those priors should take.

As one of the major goals of setting prior distributions on parameters is to minimize the occurrence of unreasonable estimates, the typical scenario is to place informative prior distributions on the discrimination parameter (a-parameter) and the "pseudo-guessing" parameter (c-parameter), as these are the estimates that tend to be most likely to go out of the expected range. This has been found to produce good results in several studies (Swaminathan & Gifford, 1986; Lord, 1986). Gifford and

Swaminathan (1990) advocate the use of item-specific priors (different priors for each item) when there is information available to do so.

Determining the appropriate prior is also critical, as it will affect the resulting estimate. Harwell and Baker (1991) showed that the closer the prior mean is to the actually parameter, the less effect the prior will have on the estimate. This is perfectly logical, as the effect of the prior is to "shrink" the estimate closer to the mean of the prior distribution. Several authors have made suggestions for the selection of prior distributions for each of the item parameters (Harwell & Baker, 1991; Swaminathan & Gifford, 1986; Zeng, 1997), and the reader is referred to these articles for the specific details. Harwell and Janosky (1991) showed that in small samples when the number of examinees is at least 250, the effect of the prior variance, that is, the amount of information contained in the prior, has minimal effect on the resulting estimates. Similarly, Gifford and Swaminathan (1990) showed that different specifications of the prior distribution had modest effects on the resulting estimates, except in cases where the distribution was extreme in nature. While these results seem to indicate that the prior has little effect on estimation, in the case of small samples the situation becomes more critical. Harwell and Janosky (1991) found that in cases of small samples (less than 250 examinees) and short tests (fewer than 25 items), the prior variance was important in obtaining proper estimates of discrimination. In a CAT context, such small tests and small samples are not unreasonable to expect, and hence, attention should be paid to the prior variance. Additionally, Seong (1990) showed that when the prior distribution for θ did not match the actual underlying distribution, item difficulties and discriminations were poorly estimated in small samples. Therefore, in

the case of small samples, greater attention to the prior distribution is necessary to obtain appropriate estimates.

2.4 Small Sample Estimation

The problem of small sample estimation is not a new one; however, the implementation of CAT makes this area of research increasingly important. Out of concern for test security and item exposure, the need to calibrate new items and replenish item banks will depend on small sample estimation techniques. There is very little literature surrounding this topic, and the methods employed are few. Among the most popular approaches are modified IRT models, optimal sampling techniques, and the use of auxiliary information in parameter estimation. Each of these topics will be discussed briefly.

2.4.1 Modified IRT Model

Research has been conducted concerning the utility of modified item response models. In modified models, models with several parameters are used (either 2 or 3), however the values of one or more of these parameters is either fixed at a certain value, or constrained to a narrow range of values. The hope is that in constraining the more complex models, the estimation process is simplified (by limiting the number of unknown parameters) and hence, the necessary sample sizes are reduced without having to use a simpler model. Most of the research has focused on constraining a 2-PL (Sireci, 1992; Stone & Lane, 1991; Harwell & Janosky, 1991; Patsula & Pashley,

1996), however two studies investigated a constrained 3-PL (Barnes & Wise, 1991; Parshall et al., 1996), which is of interest here.

The results of the Parshall et al. study (1996) concluded that by constraining the IRT models, estimates of item parameters became more stable but less accurate. Neither of these results is surprising; as the model becomes constrained, there is less freedom in the estimation process, leading to more stability and less accuracy. The Barnes and Wise (1991) study, however, showed that the use of a fixed *c*-parameter lead to more accurate recovery of both item and ability parameters. Similarly, Thissen and Wainer (1982) recommended the use of a mixed-model approach where the fixed lower asymptote was used only in the case of easy items when large samples were unavailable for estimation. However, although the results of the modified IRT model approach provide some promise, they are not consistent across studies, and when combined with the loss in flexibility additional techniques warrant further investigation.

2.4.2 Optimal Sampling

Several studies have explored the feasibility of using optimal samples to reduce the sample size necessary for item calibration (Berger, 1991, 1992, 1994; Jones & Jin, 1994; Slater, 2001; Stocking, 1990; Timminga, 1995; Yu & Way, 1998). The primary use of optimal sampling designs is in the area of pretesting items in a CAT environment. The idea of optimal sampling is to match examinees with the items that are of appropriate difficulty. The goal is to choose examinees that will provide the most information about the items administered. With the exception of the work done

by Slater (2001), the studies above required knowledge of the item and/or ability parameters. As this situation would never exist in an operational setting, the methods proposed are of limited value at this stage.

The work of Slater (2001) is very interesting and involves the use of expert judgments on the item difficulty to identify the focused sample. By using these expert judgments and the estimates of ability based on operational items, examinees can be matched to the appropriate items in the pretesting stage. Furthermore, the ability parameter can be assumed know in the estimation process, reducing the item estimation phase to logistic regression. The results of the Slater (2001) study indicate that focused samples of examinees performed well in the case of extreme values of difficulty, however, she concludes that "the results of this study support the methods in use and do not suggest devoting time and resources…for the purposes of matching item difficulty of pretest items with estimates of examinee ability." Furthermore, the use of item-specific priors is not recommended in the instance of focused sampling.

2.4.3 Use of Collateral Information in Estimation

Collateral, or auxiliary information about both items and examinees is often available in testing situations. Such information for items may include item type, presence of a figure/graph, number of words, average response times, and for examinees may be variables such as demographic information (age, gender), grades in courses, and courses taken. This information, while often available, is rarely used in the estimation of item parameters. There has been minimal research on incorporating this type of information into the estimation process. Mislevy & Sheehan (1989) have

considered the use of collateral information about examinee in the estimation process. In this study, this information was used in a MML context to enhance the specification of the distribution of θ. They then considered the affect of using collateral information on the consistency of the resulting parameter estimates. Their work shows that if collateral information is available and is used in examinee sampling and item assignment, then ignoring this information in the estimation will lead to inconsistent MML estimates. However, if this information is not used for sampling or assignment, then the estimates are consistent regardless of whether the information is used or not. Mislevy (1988) used collateral information on both items and examinees to enhance Bayesian estimation techniques. In this case, the specification of both the prior distribution of θ and the prior distribution of the item parameters are enhanced by this collateral information. More specifically, if $f(a,b,c)$ is the joint prior density for the item parameters without collateral information, $f(a,b,c|z)$ is the joint prior density of the item parameters with collateral information, where z is the vector of collateral information for each item. These priors are then used as above in the BME. The results of the study indicate that including this collateral information leads to modest improvements in item parameter estimation.

Swaminathan et al. (in press) took a different approach and used expert judgments about item difficulty to specify item-specific priors. Prior distributions were placed only on the item difficulty parameters, as judging the discriminating power of an item is an unreasonable task. Specialists were trained to estimate the difficulty of each item in terms of the proportion of examinees the raters expected to get the item right. The average of these judgments was then transformed to the scale of the IRT

difficulty parameter, and was used as the mean of a normal prior distribution. The standard deviation of the distribution was varied. The results of this study indicate that incorporating judgmental information about the difficulty of an item lead to dramatic improvements in the estimation of the *a*- and *c*-parameters. Given the substantial improvements in estimation for a short test (21 items) with small samples (100 to 500), this procedure shows great promise for small-sample calibration. However, obtaining subjective judgments from experts may be costly and time-intensive. Therefore, other types of information, which can be more objectively and routinely obtained, may be used in place of the expert judgments, and could lead to a more practical approach.

Additionally, Swaminathan et al (in press) showed that placing prior distributions on the difficulty parameter may lead to better estimates of the less stable discrimination and guessing parameters. As noted above, previous studies have emphasized the use of prior information on the *a*- and *c*-parameters, as they are most difficult to estimate. The results of Swaminathan et al. (in press), however, indicate that an appropriate approach may be to place priors on the difficulty parameters alone.

Clearly, the use of collateral information to specify item-specific priors seems to be an approach with some promise in the case of small sample estimation. Whether this information is judgmental or can be collected routinely as objective information, its use can lead to a decrease in the necessary sample sizes. With the growth in automatic item generation, the use of collateral information is becoming more central to the creation of items whose psychometric properties are known. This body of literature can be used to identify the types of collateral information available, and its

effectiveness in determining item characteristics such as difficulty and discrimination, which can then be incorporated in the estimation process via Bayesian techniques. Therefore, a discussion of collateral information, as used in item generation techniques is discussed next.

2.5 Collateral Information

The demand for large numbers of items to construct and maintain the large item pools necessary for CAT, as well as the desire to reduce pretesting and item calibration, has lead to research in automatic item generation (AIG), where items with known item characteristics can be produced by computer. A consequence of this research has been an investigation of item characteristics that can help predict the item parameters. This information can be used to enhance the method used by Swaminathan et al. (in press).

Not surprisingly, the majority of research focuses on the prediction of item difficulty. Embretson (2002) suggests the use of cognitive models in order to predict item difficulties. Using these cognitive models, item parameters can be predicted, and hence the items generated from the model can be banked without the need for calibration. The results of the research cited in her book indicate that the item parameters can be predicted fairly well. In work with quantitative items, the item parameters were predicted quite well, with $R^2=.90$. Additionally, according to Scheuneman, Gerrtiz and Embretson (1991), the use of structural variables, readability measures and semantic variables leads to successful prediction of item difficulty for passage-based items. For these types of items, the R^2 values range from .24 to .36,

indicating good prediction of item difficulty from factors that are somewhat easy to attain. Additionally, the items produced using the cognitive models appear to be different, despite the similar structures. Therefore, while the work with cognitive models may be limited by practicality, the success with the reading comprehension questions is promising, as well as practical.

Similar to the work of Embretson et al., Perkins (1995) consistently predicted item difficulty (R^2 =.74 to .96, depending on item set) based on text structure, propositional analysis, and cognitive demand. The types of variables of interest in the prepositional analysis include number of arguments, modifiers and predicates as well as the density of each of those components (e.g. argument density is equal to the number of arguments divided by the number of sentences). For text structure variables, information such as number of lines per passage, number of content words per page, the word to sentence ratio, and the percent of content words were used in the prediction. Therefore, the use of the tools of cognitive psychology and linguistic analysis can be used to predict item parameters with very good success. Given the ability to predict item parameters, calibration sizes may be able to be greatly reduced.

To avoid the complexity of the cognitive model approach, Dennis et al. (2002) considered identifying item features which could be used to predict item parameters, and in particular, item difficulty. Two studies were discussed, and in both situations, the item difficulty was sufficiently predicted from the identified item characteristics. Using semantic variables and linguistic tools, the prediction of item difficulty were quite impressive, with R^2 values ranging between .78 and .88. These items were from the Directions and Distances Test of the Royal Navy. The content of the items was

very specific, as is evidenced by the title of the test. Dennis et al. also extended this work to the GRE Analytic Reasoning items, and for items of that type, the difficulty was predicted with R^2 =.77. The prediction for items of that type relied on a classification of the options in terms of informativeness, possibility, and impossibility. While these three item types are fairly different, in all cases the item difficulty was predicted quite well from item attributes that are easily identified.

Similarly, Hornke (2002) described a series of studies conducted to the degree to which item features could predict item difficulty. Several different item types were considered in the studies: mental rotation, pattern matrices, number problems, visual analysis, visual memory, and verbal memory. In many cases, item design rules were used for prediction. Rules included features such as complexity of image, imagery, inspection time (i.e. time allotted for candidate to look at image), pattern simplicity, homogeneity, compactness, and background complexity. The specifics for each item type are detailed in the article; however, this sampling was included to indicate that the type of information used is readily available for item writers, as these are the guidelines used for item design. The correlation of the predicted values with the IRT parameters varies somewhat by item type, but the values range from .59 to .94, indicating good prediction of item parameters from item design rules. Again, while these item types are limited in scope, the evidence contributed by this study builds the bank of item types whose difficulties are predictable.

A similar approach is implemented in the Test Creation Assistant (TCA), developed by Educational Testing Service. Descriptions of the software can be found in Bennett (1999) and Singley and Bennett (2002). Basically, an item model is

developed and entered into the system. Variables within the item are identified and can be manipulated by users to produce "new" items. The software then provides a predicted difficulty for the generated item. Given the vast array of item types created by the TCA, the details of each item type are not presented here.

Enright et al. (1999) concluded that given the difficulty in explaining constructs thoroughly enough to identify the features of an item that can lead to item parameter prediction, using correlations between item features and item statistics can help identify those features which predict item difficulty. The results of the Enright et al. (1999) study were very promising, and were able to identify features that could account for 90% of the variance in difficulty. Two types of items were studied: probability problems and rate problems. Using content/context variables (e.g. percent problems, cost problems) as well as a complexity rating (2 or 3 levels, depending on type), the item difficulties were predicted with R^2= .91 for rate items and .62 for probability problems. Furthermore, these variables predicted item discrimination with an R^2 =.52 for the rate problems, however, no information regarding predicting the discrimination the probability items was provided. Again, these are limited types of items, however in both cases the item difficulty can be predicted at a fairly high level for both item types. Results regarding reading comprehension items are also prominent in the literature, and will be presented next.

Work with the Test of English as a Foreign Language (TOEFL) reading comprehension items (Freedle & Kostin, 1993) found that the difficulty of items correlated with text-related variables with ρ=.60. Freedle and Kostin also did similar analyses with the SAT and GRE (1991, 1995) and found structural (e.g. number of

sentences, length of longest paragraph) and cognitive demand variables (e.g. concreteness of text, rhetorical format) provided good prediction of item difficulty, with a multiple correlation ranging from .68 to .76 (depending on sample).

In addition to identifying item features that can be used to predict item parameters, some research on the correlation between response time and item difficulty is emerging. In a study by Halkitis and Jones (1996), the logarithm of the response time was correlated both with item difficulty and discrimination at levels -.43 and .31, respectively for items on a real estate exam. While response times are not easily obtained in the paper-and-pencil format, in the CAT environment this type of evidence is trivial to collect. This evidence could be especially useful as it may predict both difficulty and discrimination, making it a very useful variable. The results of this study were replicated by Mason (1992) who found that response time and difficulty were correlated at the .6 level for mathematics items. Clearly, more research involving the use of response-time data could be useful in the estimation of parameters. Very little research has been done with response times, however that may change with the growth of CAT in educational testing.

The studies discussed above all considered different item types. While each study was limited by the specificity of the items studies, taken as a whole, there is a substantial amount of evidence supporting the predictability of item difficulty across a wide range of item types. Additionally, these studies provide guidance as to what types of collateral information might be useful in predicting the difficulty of other types of items.

Some empirical studies have been conducted to investigate how well the automatically generated items perform; that is, the accuracy of the predictions (Lewis, 2001; Mislevy, Wingersky & Sheehan, 1994). Although the theoretical basis implies that items generated from models have similar item statistics, in reality, the small "cosmetic" changes may lead to drastically different item parameters. Therefore, it may be unwise to use the predicted parameters as the actual item parameters; however, these predicted item parameters could be used in the estimation process to refine the estimates, and to allow for smaller sample sizes.

2.6 Summary

The importance of proper estimation of item parameters cannot be understated. This chapter described the most popular methods for item parameter estimation as well as their feasibility in small sample situations. As the popularity of CAT increases, methods that allow for proper estimation of item parameters using small samples become increasingly important. Despite its importance, very little research has been conducted to devise practical approaches to improving small sample estimation. The previous work done in this area is presented in this chapter. Additionally, new work in automated item generation (AIG) was explored, and its utility in estimation was delineated. By combining the techniques of AIG in predicting item parameters with small sample techniques, a viable alternative for enhancing item parameter estimation can be achieved. The precise methodology used and the results of obtained are detailed in the following chapters.

Chapter 3

METHODOLOGY

As the purpose of this study is to explore the potential for using collateral information on the accuracy of parameter estimation, a simulation study will be conducted, since only in a simulation study can the true item parameters be known. While the context provides motivation for the need for small sample calibration, the study presented below is non-adaptive. A representative group of examinees is presented with a set of items to which they respond non-adaptively. Therefore, while the technique presented here would be useful to building a large calibrated item pool that could be used in a CAT environment, it is not limited to that context. This chapter is presented in three parts: simulation conditions, procedure and data analysis.

3.1 Simulation Conditions

Data corresponding to 12 different conditions were simulated based on different numbers of examinees and number of items administered. The specifics of each factor are provided below.

3.1.1 Sample Size

Because the purpose of the study is to investigate the feasibility in calibrating items using small samples, three small sample sizes are considered: 100, 200 and 500 examinees. One hundred examinees is an extremely small sample, while 500

examinees may be considered moderately reasonable. By investigating all three small samples, the critical number of examinees necessary to produce adequate estimates can be evaluated. Additionally, an adequate sample size of 1000 examinees is also included as a basis of comparison.

3.1.2 Number of Items

In the study conducted here, the ability of the examinee is assumed unknown. Given that fact, the number of items administered to an examinee will affect the estimation of the item parameters. Therefore, three different sets of items are administered to the examinees. A small number of items are chosen again to try to gauge the limits of the procedure as well as a more reasonable number of items. In this case, 15, 25, and 40 items were administered to all examinees.

Each of the twelve data sets that result from a cross of the two conditions above are calibrated using 10 different prior distributions. The specification of these prior distributions follows below in the procedure section of this chapter.

3.2 Procedure

In this section, the specific steps that were used to simulate the data and determine the prior distributions of the parameters are described in detail.

3.2.1 Step 1 – Generating Item Parameters and Collateral Information

Following the lead of Enright et al. (in press), the collateral information used in this study can be thought of as item features that correlate with the item parameters. In

a simulation situation, these features do not need to be specified; all that is required is data that correlates with the item parameters at the specified level. In order to obtain that type of data, the following relationship between item parameters (b_i) and the predictor variable (x_1) can be used:

$$b_i = \rho x_1 + e_1\sqrt{1-\rho^2}$$

where b_i is the $b-$parameter, ρ is the correlation coefficient, x_1 is a generated standard normal variable (i.e. $x_1 \sim N(0,1)$), and $e_1 \sim N(0,1)$ and represents random error

In this way, the *b*-parameters are generated along with the collateral information that will be used for predicting the item parameters.

Since the quality of collateral information that exists for items is certain to vary from situation to situation, the strength of the correlation with the item parameters is varied in the simulation process above. Two levels of correlation are considered: .4, and .6, indicating slightly more information about the items, while still remaining realistic. In both instances, the *b*-parameters remained constant, and the *x* values varied according to the correlation.

Since there is little evidence regarding the feasibility of predicting *a*-parameters, they were generated according to a uniform distribution in the interval [.4 , 2.] These values reflect the range of *a*-values typical in many large-scale testing programs. Similarly, the *c*-parameters were also generated from a uniform distribution: $U(0.0, 0.25)$.

3.2.2 Step 2 –Simulating Examinees

To generate item responses for each condition, a group of 1,000 examinees were simulated by generating ability parameters will from a standard normal distribution: i.e. $\theta_j \sim N(0,1)$ where j specifies an examinee.

3.2.3 Step 3 – Generating Item Reponses

Once the true item and ability parameters are known, item response data can be generated for examinees. Using both the item and ability parameters, item responses for each simulated examinee were generated. Since the form of the item response model, as well as all parameters, are known, the probability of an examinee (with the given ability) answers a particular item (with the given parameters) correctly can be calculated. Once the probability of a correct response is known, this probability is compared with a number randomly chosen from the interval (0,1). If the probability is greater than the random number, the examinee received a correct response to the item otherwise the response was incorrect. In this manner item responses for all examinees to all items were obtained, and these data can be used to estimate item parameters.

3.2.4 Step 4 – Predicting the *b*-parameters

In order to specify the prior distributions for the *b*-parameters, the item parameters need to be predicted from the collateral information generated in Step one. The approach taken here is to predict item parameters using the information available from items in the bank. Using items with known parameters along with their values for the collateral information, a multiple regression can be performed with the collateral

information for all items as the independent variables and the item parameter as the dependent variable. The predictor for *b*-parameters will be used in the prediction equation to predict the *b*-parameters, which will then be used in the estimation process. In practice collecting collateral information for items whose parameters are known will allow for the prediction of item parameters for items with unknown parameters. The specifics of the regression follow:

$$E(b) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_i x_i$$

where b is the b-parameter to be predicted, $x = (x_1 ... x_i)'$ is the vector of predictors for item parameters , and $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, ..., \beta_i)'$ is the vector of regression coefficients (in the case of the simulated data, i=2)

The regression coefficients can then be estimated given the observed values of the item parameters and the values for the collateral information as follows:

$$\hat{\boldsymbol{\beta}} = (X'X)^{-1} X'\boldsymbol{b}$$

where $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, ..., \hat{\beta}_i)$ is the vector of estimated regression coefficients, X is the matrix of collateral information, and $\boldsymbol{b}$ is the vector of known b-parameters

Given the prediction equation and items with unknown parameters, but known collateral information, the item parameters can be predicted. More specifically,

$$\hat{b} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + ... \hat{\beta}_i x_i$$

where $\hat{b}$ is the estimated b-parameter, $\boldsymbol{x}$ and $\hat{\boldsymbol{\beta}}$ are as above.

Additionally, the variance of the estimated regression coefficients can be obtained as follows:

$$V(\hat{\boldsymbol{\beta}}) = s^2 (X'X)^{-1}$$

where s^2 is MSE, and X is defined as above.

This variance will be used to generate the variance of the prior distributions.

3.2.5 Step 5- Specifying the Prior Distributions

The choice of prior distributions placed on the parameters will affect the quality of the resulting estimates, especially in the small sample case. Therefore, different combinations of prior distributions were placed on the *a*- and *b*- parameters. The prior distribution for the *c*-parameter was not changed in this study, however. As Harwell and Janosky (1991) noted, the prior variance greatly affected the estimates of item parameters, especially in small samples and short tests. Further, if the variance is too small, then it will have the effect of fixing the item parameters at the mean of the prior distribution, while a variance that is too large will have little effect on the estimation process. Furthermore, more informative priors often lead to more biased estimates, while less informative priors lead to less biased estimates. Therefore, prior variance is manipulated for both the *a*- and *b*-parameters.

Two prior distributions were chosen for the *a*-parameter. In both cases, the prior was a normal distribution for the log of the *a*-values. The mean of the prior distribution was the same in both instances (zero), while the variance of the distribution changed. In the first case, the default variance of .5 was used, while in the second case, a variance of 1 was chosen. A larger variance was chosen so as to try to decrease the bias of the estimates of the *a*-parameters, which is a primary goal of the study.

In specifying the priors for the *b*-parameters, five prior distributions were chosen. In all cases, a normal distribution was used, and the means and variances of

the distribution were changed. The first case corresponds to the default prior in BILOG, which consists of a mean of zero and a variance of 2. For the item-specific priors, the means of the distributions were the predicted parameter values. Therefore, a predicted *b*-parameter was obtained for each of the two correlations (using the corresponding set of collateral information). From each of these two regressions, a prior variance was obtained as described above. This variance was used as the prior variance in the appropriate distribution. Again, to consider the affect of the prior variance on the resulting estimates, twice this variance was also used. Therefore, four prior distributions resulted from the prediction of the parameter.

The prior on the *c*-parameter was the default prior in BILOG, and was a Beta (6, 16) distribution.

3.2.6 Step 6 – Estimating the Parameters

Once the data were generated, and the prior distributions were specified, the data were calibrated using BILOG 3. The item-specific priors were input into the BILOG program. Bayes modal estimates were then obtained for each of the three item parameters.

3.3 Data Analysis

As the purpose of the study was to improve the estimation of item parameters as well as the estimation of item information functions, the data were analyzed in two parts. First, the analyses used to examine the accuracy of the parameter estimation are

described, followed by a description of the analyses used to ascertain the recovery of the item information functions.

3.3.1 Item Parameter Recovery

Since the item parameters are used to select items in the computer adaptive environment, the accuracy of the parameter estimates is of concern. Therefore, the accuracy of the estimation procedure will be evaluated by considering the root mean squared error between the estimate (t) and the true value (τ):

$$RMSE = \sqrt{\frac{\sum_{r=1}^{R}(t_r - \tau)^2}{R}}$$

where r is the replication, and R is the number of replications.

Since MSE=(Bias)^2+Variance, the effect of bias and variance on the MSE can be evaluated by computing:

(1) The variance over replications:

$$Var = \frac{\sum_{r=1}^{R}(t_r - \bar{t})^2}{R}$$

t_r is the estimate of the parameter in replication r, R is as above, and $\bar{t}$ is the mean of the estimates across replications, and

(2) The bias:

$$Bias = \bar{t} - \tau$$

where $\bar{t}$ is the average estimate across replications, and τ is the true parameter.

Since determining what constitutes a practical decrease in MSE, and hence bias and variance, is difficult, any estimation method that produces a decrease is worthy of consideration, particularly if the proposed method does not require any additional cost or time investment. As the method proposed here can be implemented easily in an operational setting, using existing software and information, a decrease in MSE would render this procedure useful given the importance of properly estimated item parameters. Additionally, an indication of more stable estimates allows more confidence in any estimates obtained. In practice, replications are not possible, and decisions about item characteristics are based on one replication. If a given estimator is not stable, then little confidence can be placed in the estimate. Therefore, demonstrating increased stability of estimation leads to more confidence in any one estimate.

3.3.2 Item Information Recovery

As one the goals of the study is to increase the accuracy with which the item information function can be recovered, the true and estimated item information functions were compared at several points along the theta scale. This is particularly important in CAT since the sequence of item administration is often dictated by the information function. The information function for the three-parameter model is given by:

$$I_i(\theta) = \frac{2.89a_i^2(1-c_i)}{[c_i + \exp(1.7a_i(\theta - b_i))][1 + \exp(-1.7a_i(\theta - b_i))]^2}$$

As most of the information in this case is concentrated on the interval (-1, 1), this interval was chosen for comparison, and the information functions were calculated at intervals of .5. Therefore, six points were used as a basis of comparison. By considering the various points along the theta scale, not only can overall differences be assessed, but also any interaction between error and level of theta.

In order to compare the recovery of each of the calibrations, the RMSE between the true and estimated information function was computed for each item and averaged across items. Given the large numbers of items and conditions, for the ease of presentation, the average RMSE over items was reported rather than the RMSE at the item level. As noted earlier, there is a problem of interpreting the RMSE in an absolute sense; however, since the RMSE is compared across conditions, and the procedure is relatively easy to implement, the condition that yields the smallest RMSE will indicate the best procedure to adopt in practice.

While an accurate recovery of the information is desirable, perhaps more important is the bias of the information function, as the information function is used as a primary criterion in item selection. Therefore, the bias of the estimated information functions at each of the six points on the theta scale was also examined.

CHAPTER 4

RESULTS

This chapter presents the results of the study. First, the parameter recovery of the estimation techniques will be presented. Parameter recovery was assessed by calculating the root mean square error (RMSE) between the true and estimated parameters, the bias and the standard deviation of the estimates. For ease of presentation, these values were averaged across items for each condition. The RMSEs are presented first, followed by an examination of the components: bias and variance. Of particular interest to this study is the estimation of the *a*-parameter. Therefore, the results for the *a*-parameter will be presented first. The effect of sample size and prior distribution will be presented for a given number of items first, and then trends across test lengths will be considered. Further, as the *c*-parameter is not used in item selection, and since there was little or no improvement in estimation, the results of the *c*-parameter will not be presented but can be found in Tables A.1 to A.3 in the Appendix. Following the results of the item parameter recovery, the results of the recovery of the item information functions are presented. Therefore, despite the lack of explicit results for the *c*-parameter, the precision of its estimation will affect the recovery of the item information function. By looking at the recovery of the information functions, the combined effect of the item parameters is considered.

In calculating the average values, item level values were summed, and divided by the total number of items. As a result, the relationship $MSE = (Bias)^2 + Variance$ is not maintained at the average level. Additionally, since bias can be both positive

and negative, the average absolute bias was calculated so that the positive and negative values did not cancel each other out. As several different prior variances are considered in the study, a few comments regarding the sizes of the variances is warranted. In all instances, the variances of the *b* priors are based on the variance obtained from the regression. In the case where ρ=.40 , the variance obtained from the regression was approximately .77, while in the case where ρ=.60, that variance is approximately .58. Therefore, the case of twice the variance leads to variances of 1.5 and 1.2, respectively.

4.1 Item Parameter Recovery

4.1.1 RMSE of the *a*- and *b*-parameters

The results for the average RMSE of the *a*-parameter for all conditions can be found in table 4.1. Regardless of the item-specific prior, the RMSE is smaller than that when the default prior is used. Similarly, for all priors, including the default prior, and all test lengths, as the sample size increases, the RMSE decreases, as would be expected. The prior that produces the smallest RMSE varies somewhat by condition; with a clear tend across test lengths.

Considering the 15- item "test" first, for all sample sizes except the 100 examinee sample size, the smallest RMSEs occur when a more informative prior is placed on the *a*-parameter, and the lower correlation is used to produce the more informative prior for the *b*-parameter. The improvements in this case range from between 8 and 14 percent decrease in RMSE. In the case of the 100 examinees, a less

informative prior on the *a*-parameter, along with the most informative prior on the *b*-parameter (corresponding to a $\rho = .60$ and the smallest variance) produces the most accurate results, and produces an RMSE that is 29% smaller than that obtained when the default prior is placed on the *b*-parameter (with the same *a*-prior). Therefore, in the most extreme case, the most informative priors produce the best results. However, as sample size increases, using less informative priors is more effective.

When the number of items increases to 25, using the more informative prior on the *a*-parameter produces the best results across all sample sizes. However, as in the previous instance, the prior on the *b*-parameter that produces the best results is common for all sample sizes except the smallest (N=100). Unlike the case of 15 items, a more informative prior on the *b*-parameter is preferred. Using the stronger correlation ($\rho = .60$) and the smaller variance results in the smallest RMSEs for the *a*-parameter. The improvements over the default prior on the *b*-parameter range from 16 to 25 percent. In the case of 100 examinees, the informative prior that results from the lower correlation ($\rho = .40$) is preferred, and leads to a decrease of 12 percent in the RMSE.

Increasing the items to 40 leads to a more consistent pattern. In all sample sizes, the same prior produces the smallest RMSEs for the *a*-parameter. As in the other cases, the more informative prior for the *a*-parameter produces the best results, along with the most informative prior on the *b*-parameter ($\rho = .60$, smaller variance). In this situation, the improvements are largest with a decrease of RMSE between 21% and 31%. Therefore, in the case of 40 items, the most effective prior is clear.

In considering the results of all the test lengths, some overall comments can be made. In general, a more informative prior should be used on the *a*-parameter. If a small number of items are used, then the smaller correlation produces more accurate estimates, and as the number of items increases, the larger correlation is preferred. In all cases, the more informative prior on the *b*-parameters leads to smaller errors in estimation.

The results for the *b*-parameter are clear, and are provided in Table 4.2. For most conditions, the same prior distribution produces the most accurate results. In this case, unlike the case of the *a*-parameter, the estimates of the *b*-parameter are most accurate when a less informative prior is placed on the *a*-parameter. Not surprisingly, the most informative prior for the *b*-parameter results in the smallest RMSEs. That is, the prior that results from the higher quality collateral information ($\rho = .60$) and the smallest prior variance lead to the most accurate estimates of the *b*-parameter. There are three exceptions: N=200, n=15; N=100, n=25; and N=100, n=40. In these instances, the more informative prior should be placed on the *a*-parameter. For the first two cases, the same informative prior should be used for the *b*-parameters, however, for the third case (N=100, n=40), the larger variance should be used in conjunction with the stronger correlation. For the *b*-parameter, the improvements in RMSE are similar to those for the *a*-parameter. The RMSEs are reduced between 7% and 20%. In this instance, the larger improvements are found in the smaller sample sizes, regardless of test length.

Again, as in the case of the *a*-parameters, an interesting result emerges when the sample size is taken into account. For the *b*-parameter, with 15 items and 500

examinees, the estimate obtained using the item-specific prior is more accurate than that obtained with the default *b*-prior and the same *a*-prior (variance equal to one). However, in the case of 40 items, the best estimate based on 100 examinees is more accurate than the estimate based on 1000 examinees, using the default prior on both the parameters.

While improving the accuracy of estimating the item parameters is of interest, more importantly, perhaps, is decreasing the bias of the estimate. An apparently highly discriminating item may in fact be a poorly estimated item with an overestimated value of the *a*-parameter. Since items with high *a*-parameter values are typically chosen in CAT, this choice can lead to more error in ability estimates. Therefore, and examination of the bias of item parameters is important and considered next.

As mentioned earlier, bias can be either positive (indicating an overestimate) or negative (indicating an underestimate), or even zero (indicating perfect estimation). It is for this reason the average absolute bias is considered, since the positive and negative biases will not cancel each other out and give the false impression of perfect estimation.

4.1.2 Bias in the *a*- and *b*-parameters

As in the case of the RMSE, using any item-specific prior produces estimates of the *a*- and *b*- parameters that are less biased than those obtained using the global default priors. The results for the average absolute bias of the *a*-parameter are found in Table 4.3. In general, regardless of condition, the less informative prior on the *a*-

parameter produces the less biased estimates. This result is not surprising, and is what would be expected. The one case where the more informative prior on the *a*-parameter is preferred is when there are 1000 examinees and 25 items. The bias in this case, .028, is smaller than that which results with a less informative prior, .038. However, the difference, .01, may not be meaningful. When the effect of the prior for the *b*-parameter is considered, some differences were observed among the conditions.

In the case of 15 items, when the number of examinees is greater than 100, the smaller correlation and smaller variance produces the prior which is most effective. However, when the sample gets large (N=1000) the default prior on the *b*-parameter produces the least biased estimates. Again, this result is not surprising. As in the case of the RMSE, for the small sample size, the stronger correlation for the *b*-prior is necessary to get the least biased estimates. However, the results of using $\rho = .60$ over $\rho = .40$ are very similar (.087 vs. .092). The improvements in bias are more dramatic than those for the RMSE. The percent decrease for the 15 item test ranges between 23% and 46%, with the largest improvements in the smaller sample sizes. That is, as the sample size increases, the improvement decreases.

As the number of items increases to 25 items, a similar pattern is observed. As mentioned above, for all sample sizes except 1000, the less informative prior on the *a*-parameter produces the less biased estimates. In the case of 1000 examinees, the more informative prior produces less biased estimates. The effect of the various priors on the *b*-parameter is similar to the previous case. In the smallest sample (N=100), the estimates are least biased when the most informative prior is placed on the *b*-parameter ($\rho = .60$, and smaller variance). However, as above, these results are very

similar to those obtained when $\rho = .40$ and the smaller variance are used (.086 vs. .094). For the remaining small sample sizes (N=200, 500), using the informative prior based on $\rho = .40$ results in the least biased estimates. When the largest sample size is considered, the less informative prior that results when $\rho = .60$ produces the least bias, although the bias is similar to that obtained when $\rho = .40$ and the informative prior is used for the *b*-parameter (.028 vs. .038). The decrease in bias for the longer test is even greater, with a 56 to 78 percent decrease in bias depending on sample size. The improvements in the case of 40 items are quite similar and will be considered next.

As with the RMSE, the pattern of results in the 40-item case is clearer. Again, as indicated above, the less informative prior on the *a*-parameter leads to the less biased estimates. In considering the various priors on the *b*-parameter, the *a*-parameter is least biased when the strong correlation ($\rho = .60$) is used and the prior variance is small. The one exception is in the large-sample case (N=1000) where the smaller correlation and the small prior variance yield the best results. The decrease in bias in this instance ranges from 55% to 75%, with the improvement increasing as the sample size decreases.

The results of the bias analyses indicate that using item-specific priors can drastically decrease the bias in the estimates. While the trends in improvement are not strictly equivalent in all conditions, some general recommendations can be made. In specifying the prior for the *a*-parameter, a less informative prior leads to less biased estimates. In choosing a prior for the *b*-parameters, the sample size and test length must be considered. For samples of 500 examinees or less, and between 15-25 items, the prior produced by the smaller correlation along with the smaller prior variance

would be preferred, while with larger samples, the stronger correlation along with a larger prior variance would be recommended. As the number of items increases, the smaller prior variance produces the least biased estimates. However, in the smaller samples (500 or fewer) the stronger correlation ($\rho = .60$) is best, while with a larger sample the smaller correlation ($\rho = .40$) is recommended. The expected decrease in bias is between 25% and 50% for smaller numbers of items and 50% to 75% with a greater number of items. Interestingly, when the *a*-parameter is considered, the prior distributions that produce the most accurate estimates (in terms of RMSE) are not always the same ones that produce the least biased estimates.

Similar to the *a*-parameter, the estimates of the *b*-parameters are less biased when item-specific priors are placed on the *b*-parameter. The results of the average absolute bias are presented in Table 4.4. The effect of the prior for the *a*-parameter varies by sample size; when very small (N=100,200) samples are used, a more informative prior on the *a*-parameter produces less biased estimates of the *b*-parameter, while when larger samples are considered (N=500, 1000), the less informative *a*-prior yields less biased results for the *b*-parameter. The choice of prior for the *b*-parameter depends on the sample size and test length, and the various test lengths will be considered next.

For 15 items, and all sample sizes, the stronger correlation produces the less biased estimates. Some differences exist in the specification of the prior variance. In all cases except when N=100, the more informative prior yields the least biased estimates. Additionally, as mentioned above differences also exist in the specification of the prior for the *a*-parameter. As noted, in the case of 100 or 200 examinees, a

more informative prior is better, while in the case of 500 or 1000 examinees, a less informative prior is superior. The reduction of bias in the estimates of the *b*-parameters is not as dramatic as that of the *a*-parameters, however, it still nontrivial. The decrease in bias ranges from 11% to 49%. The largest improvement is when N=200. The trends in improvement are almost identical for 25 and 40 items, and as such will be discussed together next.

When 25 or 40 items are administered, the specification of the prior for the *b*-parameter that yields the least biased estimates varies by sample size. In the smaller sample cases (N=100, 200) the stronger correlation with the larger variance produces the best estimates, while in the larger sample sizes (N=500, 1000), the smaller correlation with the smaller variance produces a superior prior. The decrease in bias is more modest in this case, with a percent decrease ranging between 9% and 33%, for 25 items and between 9% and 36% for 40 items, with the improvement decreasing as sample size increases.

As has been indicated by the results presented above, the use of item-specific priors can greatly reduce the bias in the estimates of the *b*-parameter. The improvement is largest when the number of items is small, and decreases somewhat when the number of items increases. The improvements are also greatest when the sample size is small. In those instances where the improvements are largest, using the less informative prior based on more (or better) collateral information is required. In the cases where less improvement is made, the priors are more informative and based on less collateral information. Similar to the *a*-parameter, the priors that yield the

smallest RMSEs are not generally the same ones that produce the least biased estimates.

While bias may be of primary interest for the purposes of this study, the standard deviations of the estimates are also important as they provide information regarding the stability of the estimates. For this reason, the standard deviation of the estimates is examined next.

4.1.3 Standard Deviation of the *a*- and *b*-parameters

In practice only one estimate is obtained for each item parameter. Given this fact, the stability of that estimate is important. If the resulting estimate is not very stable, then its value is limited since a different sample of examinees would result in a different estimate. In the context of a simulation study, however, the stability of an estimate can be evaluated by considering the variance, or standard deviation, of the estimates across replications. The standard deviations of the estimates of the *a*-parameter are provided in Table 4.5. Since there is very little difference between the conditions, all test lengths will be considered simultaneously.

Regardless of sample size or test length, the more informative prior on the *a*-parameter produces the most stable estimates of the *a*-parameter. This is not surprising, and is what would be predicted. In selecting a prior for the *b*-parameter, the prior that produces the most stable estimates of the *a*-parameter varies slightly. In all but two cases, the most informative *b*-prior produces the best results. That is, the prior based on the larger correlation ($\rho = .60$) along with the smaller prior variance produces the most stable estimates of the *a*-parameter. The two exceptions are in the

case of 15 items, and 100 or 200 examinees. In these instances, the informative prior based on the smaller correlation ($\rho = .40$) produces more stable estimates.

Although the specific priors that lead to the most stable estimates of the *a*-parameter do not vary with test length, the amount of improvement does. The shortest test length produces the least improvements, with a decrease in variability of 10% to 25% over the default priors. As the sample size increases, so does the improvement. As the number of items increases to 25, so does the improvement. In this instance, the percent decrease is between 13% and 40%, again with the improvements increasing with sample size. In the longest test, 40 items, the improvement stabilizes, with improvements between 20% and 36%. As in the other cases, the improvements increase with sample size. Therefore, in the small sample cases, the improvement in variability of the estimates is minor. It should be noted that the priors that produce the smallest RMSEs also produce the most stable estimates. The discussion of the item parameter recovery ends with a discussion of the stability of the *b*-parameter estimates. These results follow.

The estimates of the *b*-parameters are also more stable when an item-specific prior is used, rather than a global prior. The results for the *b*-parameter are presented in Table 4.6. Similar to the *a*-parameter, the results of the *b*-parameter are consistent across test lengths. As expected, the more informative prior distributions based on $\rho = .60$ are superior to those based on $\rho = .40$. There is little or no pattern to whether the *a*-prior should be more or less informative. However, in most cases, there is little difference in the results depending on the *a*-prior. The one exception is when 200 examinees are administered 25 items, in which case, the more informative prior is

superior. In terms of improvement in variability over the default *b*-prior in all test lengths the improvement is approximately 15% to 20%, with little difference between sample sizes. Therefore, the same improvement can be obtained regardless of test length and sample size. Unlike the results of the *a*-parameter, however, the congruence between the priors that produce smaller RMSEs and the priors that produce less variability is only about 50%. In the remaining cases, the priors that produce less biased estimates are the same as those that produce more stable estimates.

While there is no one prior distribution that improves the estimation of both the *a*- and *b*-parameters on all criteria of interest (RMSE, bias, SD), it is clear that using item-specific priors improves the estimation of both parameters. However, it is not necessary for all parameters to be affected equally by a particular prior distribution, as individual parameters are rarely of interest. What is more important is the combination and interaction of the item parameters. In this study, the motivation is to improve item parameter estimation so as to improve item selection, and hence ability estimation. Since item selection often depends on the item information functions, the criterion of interest here is the recovery of the true item information functions. By considering the recovery of item information functions, the combined effects of the item parameter estimates can be considered. Therefore, for each method of estimation, the item information is compared to the item information based on the true item parameters at several points along the ability scale. The results of the recovery of the information function are presented in the next section.

4.2 Recovery of Item Information Functions

As mentioned above, the deviation between the estimated information function and the true information function was evaluated for all conditions at various points along the ability distribution. Item level results were not practical to present due to the large number of conditions and items per condition; hence, summary statistics were required. The RMSE between the true and estimated information functions was calculated at 13 points on the theta scale, ranging from –3.0 to 3.0. As the b-parameters were simulated to be distributed approximately *N(0,1)*, there is likely to be very little information at the tails of the theta distribution, since the peak of the information function occurs at the point on the theta scale equal to the b-value of the item. Not surprisingly, the information function was recovered equally well at the tails of the distribution across all conditions, and hence the results presented focus on theta values between -1.0 and 1.0, at intervals of .5, where the differences are greatest. Furthermore, it is this interval where the majority of the examinee population exists in most cases. Additionally, as bias is of primary concern in this study, the bias between the true and estimated item information functions was calculated along the theta scale.

The results of the information recovery will be presented in two parts. First, the RMSE between the true and estimated information functions will be presented for each condition. Second, the bias of the estimated information functions will be presented. Within each part, the results will be presented by test length, to be consistent with the item parameter recovery section. The section will conclude with some general comments on the recovery item information functions.

4.2.1 RMSE of Information Functions

As noted, the results are discussed according to test length. First, the recovery of the item information functions for 15 items will be presented. For each sample size, there are two graphs, corresponding to the two different prior variances on the *a*-parameter. These curves were not placed on the same graph for ease of reading; however, the scale was maintained within sample size to aid in comparison of the two figures. The 100- and 200-examinee sample size will be considered together, as the trends are the same. The results are presented in Figure 4.1 and 4.2 for 100 and 200 examinees, respectively.

It is clear that when the less informative *a*-prior is used, the informative item-specific priors on the b-parameters for both correlations produce the best recovery of the information functions. When the variance of the *a*-prior is increased, however, the two priors corresponding to $\rho = .60$ outperformed the other priors, producing the most accurate information functions. The prior that leads to the best recovery of the item information corresponds to the case where the less informative prior is used for the *a*-parameter, and the less informative prior obtained using $\rho = .60$ on the *b*-parameter. This prior distribution leads to a decrease in RMSE of 29% to 52% in the case where N=100 and 18% to 38% when N=200 over the default prior.

As the sample size increases to 500, the RMSE for all priors is greatly reduced indicating better recovery of the information function regardless of prior distribution. Figure 4.3 provides the results for the 500-examinee case. In terms of improvement in recovery, a pattern similar to that of the other sample sizes is observed. As in the other two sample sizes, in the case of the more informative *a*-prior, the two more

informative *b*-priors (corresponding to $\rho = .40$ and $\rho = .60$) are generally better, and as the prior variance increases for the *a*-parameter, the two priors resulting from $\rho = .60$ recover the information functions the best. Again, in terms of best recovery, the less informative prior on the *a*-parameter leads to better recovery. Depending on the point on the theta scale that is of interest, either the more informative prior corresponding to $\rho = .60$ or the less informative prior based on is $\rho = .40$ is preferred, although the RMSE values are quite similar for both priors. Therefore, for the sample size of 500, the prior variance of the *b*-parameter is of less importance, provided it is based on collateral information. The RMSE for all priors is relatively small at this sample size; the improvements over the default are relative smaller than that observed sample size of 100 and 200 with the RMSE showing a 13% to 27% decrease over the default priors.

For 1000 examinees, the RMSEs for all prior distributions are small, and are very similar to the 500-examinee case, indicating that the recovery of item information is good with as few as 500 examinees. The RMSEs for the various conditions in this case are provided in Figure 4.4. Clearly, despite the small RMSEs in the default case, there are improvements, even in the case of 1000 examinees. In this instance to two priors that lead to the best recovery of item information are based on $\rho = .60$, regardless of the prior on the *a*-parameter. However, as before, the less informative *a*-prior leads to the best item information recovery coupled with the more informative prior resulting from $\rho = .60$ on the *b*-parameter. The decrease in RMSE in this instance was even smaller, as a result of the better recovery for all priors, however improvements of 7% to 15% were obtained using the item-specific priors.

Given the results of the 15-item test for all sample sizes, some general patterns have emerged. Overall, the best recovery of the item information functions occurs when a less informative prior is placed on the *a*-parameter, and the best collateral information is used for the prior on the *b*-parameter. There were some differences in determining the best prior variance to use for the prior on the *b*-parameter, however the differences in results between the two prior variances were in general small, except in the smallest sample case.

Increasing the number of items administered to 25 leads to some changes in the recovery of the item information functions; however, some of the same general patterns remain. Again, starting with the small sample size (N=100), there is a change in the trend for the *a*-parameter. The results for the parameter recovery for this sample size are provided in Figure 4.5. In this instance, the more informative prior for the *a*-parameter leads to better recovery of item information. The more informative prior on the *b*-parameter that results from the smaller correlation produces the best results. The decrease in RMSE ranges from 18% to 25%. It should be noted that the results of the combination of priors where the less informative prior is placed on the *a*-parameter and the less informative prior based on $\rho = .60$ is placed on the *b*-parameter.

Increasing the sample size to 200 examinees the trends from the previous test length emerge, as is evident in Figure 4.6. In general, except where there is little difference between the priors, the less informative prior on the *a*-parameter combined with the *b*-prior consisting of the higher quality collateral information and the smaller variance produce the smallest RMSEs between the true and estimated information

functions. In this case the size of the improvement is sizable with a decrease of 23% to 47% over the default prior on the *b*-parameter.

In the case of 500 examinees, there is less distinction among the item-specific priors. All item-specific priors outperform the default, global priors for the *b*-parameter, however, three of the four lead to very similar results, as is shown in Figure 4.7.

The one item-specific prior that does not lead to much improvement is the prior where the lower correlation ($\rho = .40$) is used to determine the mean, and the larger variance is used. With the other item specific priors, the point of the theta scale, to some extent, determines which prior performs best. In general, however, the less informative *a*-parameter produces the smallest RMSEs with the *b*-priors based on $\rho = .60$ producing better results. The more informative *b*-prior recovers the item information more consistently across the theta scale than the less informative prior on the *b*-parameter. Considering the more informative *b*-prior, the reduction in RMSE is between 5% and 25%.

Increasing the sample size to 1000 examinees, the RMSEs for all prior distributions is reduced. The results for the large sample case are provided in Figure 4.8. Again, as in the case of 500 examinees, the differences between the item-specific priors are small. However, it is consistent with the results observed with n=500; the less informative prior on the *a*-parameter leads to the best recovery of the item information functions. Among the item-specific priors in that case, the most informative prior on the *b*-parameter (based on $\rho = .60$ and the smaller variance) leads to the best recovery, with a decrease of 1% to 4% in RMSE over the default prior.

These improvements are small, as would be expected given the smaller RMSEs for all prior distributions, as would be expected given the large sample size.

Again, across the sample sizes some general patterns emerge. In all but the smallest sample case, the less informative prior on the *a*-parameter and the most informative *b*-prior ($\rho = .60$, smaller variance) produces the most accurate estimation of the item information function. In the smallest sample case, however, the more informative prior is needed on the *a*-parameter, as well as the *b*-prior based on the lower correlation.

Turning to the last test length, 40 items, the trends are very similar to those in the 25-item case. Inspection of Figure 4.9, with a sample size of 100, the best recovery of the information is obtained when the more informative prior is placed on the *a*-parameter for all priors on the *b*-parameter. When the more informative priors are placed on the *b*-parameters, the recovery is even better. Both the prior based on $\rho = .40$ and $\rho = .60$ produce good recovery, with the larger correlation producing the better results, in general. The reduction of RMSE when using the most informative priors is between 24% and 35%.

When 200 examinees are administered the 40 items, there is slightly better recovery of the information when the more informative prior is placed on the *a*-parameter. The results of the recovery for the 200-examinee case are presented in Figure 4.10. In this instance, the prior for the *b*-parameter that produces the best results is the one resulting from the correlation of .60 and the smaller variance. The reduction in error is between 27% and 40% over the default prior. The results are very

similar to the case where the same prior is used on the *b*-parameter, but the less informative prior is used for the *a*-parameter.

For both the 500- and 1000-examinee cases, the less informative prior for the *a*-prior produces the best recovery of the information function. Figure 4.11 displays the results for the 500-examinee case while figure 4.12 provide the results for the case of 1000 examinees. Again, the most informative prior on the *b*-parameter ($\rho = .60$ and small variance) produces the best recovery. The percent decrease in RMSE is between 12% and 34% for 500 examinees and 32% to 35% for 1000 examinees.

Examining the trends across the various sample sizes for this test length, the superior prior for the *b*-parameter is obvious; the prior based on $\rho = .60$, along with the smallest variance produces the best estimation of the item information function. As in the case of the 25-item test, the smaller sample sizes require a more informative prior on the *a*-parameter.

Unlike the results for the item-parameter recovery, the results of the item-information recovery are fairly consistent across the conditions. It is clear that when the better quality collateral information is used (represented by $\rho = .60$), the recovery of information is best. There is one exception to this case, yet the results for the case of $\rho = .60$ are very similar. In terms of the prior variance of the *b*-parameter, there are some differences, although overall the smaller variance is preferred. The only instance where the larger variance is preferred is in the 15-item case and the small sample sizes (N=100, 200). In terms of the prior on the *a*-parameter, there is a clear pattern as well. In general, the less informative prior leads to the better recovery, however there are a few cases where this is not true. When the sample size is small,

especially relative to the number of items to be calibrated, the more informative prior is required to get the accurate estimates of item information, as expected.

While the recovery of item information functions is certainly of central concern to this study, the systematic error is even more central. By considering the bias of the estimated information functions, the extent to which the error is systematic can be evaluated. Therefore, the following section presents the results of the bias of the information functions across test length and sample sizes.

4.2.2 Bias of Information Functions

As in the previous sections, the results of the bias analyses are presented by test length. As expected, regardless of test length, the bias of the information functions generally decreases with sample size. The patterns with the bias are less clear, as the prior that produces the least biased estimate changes depending on theta. Therefore, in presenting the results, either a general trend will be noted, or more specific detail regarding the appropriate points on the theta scale will be given. The presentation of results begins with the shortest test length first.

The results for 15 items and 100 examinees are presented in Figure 4.13. Throughout most of the theta scale, the bias is smallest when the less informative prior is used for the *a*-parameter. The *b*-prior which produces the smallest bias, in general is the prior based on $\rho = .60$, with the smaller variance. Where the informative prior is best, the decrease in bias is 21% to 43% over the default priors. The only exception is when $\theta = -1.0$, where the least informative item-specific prior produces the best results ($\rho = .40$, larger variance).

The results for the 200-examinee case are provided in Figure 4.14. At the lower end of the ability scale, the more informative prior on the *a*-parameter produces the least biased results, while for the upper end, the less informative *a*-prior produces less biased results. In both instances, the priors for the *b*-parameter that produce the least biased estimates result from $\rho = .60$. At the lower end, the less informative prior is superior, with a decrease in bias between 29% and 42%, while at the upper end, the more informative prior is preferred, leading to a decrease in bias of 21% to 34%.

When the sample size increases to 500 examinees, the effect of the priors on the *a*-parameters becomes clear. As can be seen in Figure 4.15, the bias is less when the less informative prior for the *a*-parameter is used, regardless of the value of θ. For θ values less than zero, the less informative prior based on $\rho = .60$ produces the best results, leading to a decrease in bias of 46% to 52%, while at the upper end, the more informative prior based on $\rho = .40$ is superior, resulting a 5% to 43% decrease in bias over the default priors.

In the case of 1000 examinees, the effects of the priors on both parameters become clear. The results for this sample size are displayed in Figure 4.16. While the bias is smaller for all priors, the improvement over the default priors occurs when the prior of the *a*-parameter is less informative. The least biased estimates then result when the prior of the *b*-parameter is based on $\rho = .60$ and has the smaller variance. The resulting decrease in bias is between 11% and 33%, with the improvement being larger for smaller values of theta.

Considering the results across test lengths, some general patterns of results are observed. In general, the less informative prior on the *a*-parameter produces the best

results along with the most informative prior on the *b*-parameter ($\rho = .60$ smaller variance). The exceptions to this trend occur for smaller values of theta. In these cases, a less informative prior on the *b*-parameter produces the least biased estimates.

The results for the 25-item case are presented next, beginning with the smallest sample size first. The results for the 100-examinee case are provided in Figure 4.17. In this case, the results are clear. The less informative *a*-prior produces the least biased estimates, especially when combined with the most informative prior on the *b*-parameter ($\rho = .60$, smaller variance). The decrease in bias in this instance is between 43% and 66%.

As the sample size increases to 200, the results are similar to the 100-examinee case. Figure 4.18 provides the results for the 200-examinee condition. The less informative prior on the *a*-parameter produces the least biased results, in general, although for some levels of theta, the differences are small. In terms of the preferred prior for the *b*-parameter the most informative prior based on $\rho = .60$ produces the best results in general, with improvements between 26% and 56% over the default prior.

Figure 4.19 provides the results of the condition of 500 examinees. The results indicate that in general the less-informative prior on the a-parameter produces the least biased results, especially when paired with the less informative prior based on $\rho = .60$ on the *b*-parameter. The decrease in bias ranges between 26% and 53% for this prior.

When the number of examinees increases to 1000, the bias in estimating the information function is small regardless of the priors used. However, some improvement is found, even in the large-sample case. The pattern is clear in this condition. The less informative *a*-prior combined with the less informative prior that

results from a correlation of .60 produces the least biased results. The decrease in bias is non-trivial and ranges between 25% and 45%, as is shown in Figure 4.20.

Summarizing the results across the various sample sizes, some general patterns become obvious. In all cases, the less informative prior on the *a*-parameter, along with the *b*-prior resulting from $\rho = .60$, produces the least biased results. The variance for the prior on the *b*-parameter depends on sample size. For the smaller samples (N=100, 200), the smaller variances provide less biased estimates, however for the larger sample sizes (N=500, 1000) the larger variance provides the less biased estimates.

The results of the 40-item test are very consistent across sample sizes. As the same pattern of results is observed for all but the largest sample size, they will be discussed together first. Figure 4.21 provides the results for 100 examinees, Figure 4.22 for 200 examinees, and Figure 4.23 for 500 examinees. In all cases the less informative prior for the *a*-parameter produces the least biased estimates and the preferred prior for the *b*-parameter results from the use of *b*-prior resulting from $\rho = .60$, and the smaller variance. The percent of improvement does vary among sample sizes. The improvement is greater for smaller sample sizes (36-70% for N=100, 30%-80% for N=200) than for the larger sample size (16-61% for N=500). When the sample size increases to 1000, the prior on the *a*-parameter is again the less informative prior, but the prior on the *b*-parameter that leads to the least biased results is the more informative prior produced by $\rho = .40$. The results for this sample size are provided in Table 4.24. The percent of improvement in this case is between 26% and 44%, depending on the value of theta.

The results are consistent across the test lengths and sample sizes, with a few exceptions. In all cases, the less informative prior on the *a*-parameter produces the least biased estimates of the item information functions. Regarding the specification of the prior for the *b*-parameter, using the prior produced by $\rho = .60$ along with the smaller variance provided the least biased results in most cases. In the larger sample (N=500, 1000) and medium test length (n=25), the larger variance outperformed the smaller variance. Additionally, in large-sample (N=1000) cases (with n=15, 40) where the prior based on $\rho = .40$ performs slightly better.

Table 4.1

Average RMSE of *a*-parameter

a Prior	ρ	*b* Prior	Sample Size 100	200	500	1000
Test Length = 15						
Var = 1	.40	Default	0.594	0.495	0.384	0.312
		Var 1	0.532	0.465	0.365	0.304
		Var 2	0.581	0.493	0.382	0.312
	.60	Var 1	0.423	0.429	0.375	0.304
		Var 2	0.457	0.421	0.357	0.308
Var = .5	.40	Default	0.472	0.435	0.383	0.343
		Var 1	0.430	0.401	0.338	0.295
		Var 2	0.464	0.429	0.369	0.319
	.60	Var 1	0.499	0.409	0.334	0.292
		Var 2	0.581	0.489	0.380	0.310
Test Length = 25						
Var = 1	.40	Default	0.602	0.489	0.354	0.291
		Var 1	0.498	0.416	0.319	0.269
		Var 2	0.570	0.469	0.346	0.285
	.60	Var 1	0.563	0.418	0.317	0.267
		Var 2	0.666	0.407	0.329	0.289
Var = .5	.40	Default	0.472	0.437	0.372	0.349
		Var 1	0.415	0.373	0.305	0.264
		Var 2	0.458	0.416	0.347	0.312
	.60	Var 1	0.421	0.366	0.298	0.260
		Var 2	0.585	0.454	0.339	0.280
Test Length =40						
Var = 1	.40	Default	0.645	0.541	0.382	0.314
		Var 1	0.526	0.453	0.331	0.277
		Var 2	0.612	0.521	0.371	0.303
	.60	Var 1	0.489	0.453	0.323	0.271
		Var 2	0.477	0.447	0.360	0.390
Var = .5	.40	Default	0.517	0.502	0.454	0.389
		Var 1	0.437	0.408	0.328	0.279
		Var 2	0.493	0.466	0.395	0.342
	.60	Var 1	0.409	0.388	0.316	0.270
		Var 2	0.573	0.504	0.356	0.291

Table 4.2

Average RMSE of *b*-parameter

a Prior	ρ	*b* Prior	Sample Size 100	200	500	1000
Test Length = 15						
Var = 1	.40	Default	0.664	0.613	0.553	0.517
		Var 1	0.621	0.576	0.523	0.495
		Var 2	0.652	0.602	0.542	0.508
	.60	Var 1	0.616	0.570	0.499	0.473
		Var 2	0.660	0.625	0.595	0.581
Var = .5	.40	Default	0.686	0.653	0.638	0.640
		Var 1	0.642	0.606	0.570	0.553
		Var 2	0.673	0.640	0.616	0.612
	.60	Var 1	0.616	0.489	0.537	0.519
		Var 2	0.638	0.590	0.530	0.496
Test Length = 25						
Var = 1	.40	Default	0.661	0.619	0.572	0.544
		Var 1	0.643	0.601	0.547	0.525
		Var 2	0.652	0.609	0.563	0.535
	.60	Var 1	0.629	0.586	0.538	0.514
		Var 2	0.623	0.633	0.619	0.616
Var = .5	.40	Default	0.682	0.660	0.662	0.675
		Var 1	0.648	0.615	0.590	0.578
		Var 2	0.672	0.646	0.640	0.643
	.60	Var 1	0.619	0.621	0.588	0.575
		Var 2	0.609	0.598	0.552	0.525
Test Length =40						
Var = 1	.40	Default	0.681	0.635	0.581	0.555
		Var 1	0.666	0.606	0.541	0.517
		Var 2	0.669	0.623	0.570	0.542
	.60	Var 1	0.641	0.593	0.538	0.512
		Var 2	0.637	0.643	0.619	0.614
Var = .5	.40	Default	0.700	0.679	0.678	0.678
		Var 1	0.658	0.619	0.587	0.577
		Var 2	0.686	0.659	0.650	0.649
	.60	Var 1	0.634	0.606	0.578	0.565
		Var 2	0.619	0.611	0.553	0.524

Table 4.3

Average Absolute Bias of *a*-parameter

a Prior	ρ	*b* Prior	Sample Size 100	200	500	1000
Test Length = 15						
Var = 1	.40	Default	0.164	0.113	0.053	0.031
		Var 1	0.092	0.072	0.041	0.042
		Var 2	0.151	0.108	0.050	0.031
	.60	Var 1	0.087	0.126	0.054	0.047
		Var 2	0.138	0.108	0.064	0.046
Var = .5	.40	Default	0.154	0.125	0.085	0.056
		Var 1	0.125	0.095	0.057	0.049
		Var 2	0.146	0.117	0.072	0.048
	.60	Var 1	0.087	0.102	0.059	0.054
		Var 2	0.143	0.099	0.047	0.031
Test Length = 25						
Var = 1	.40	Default	0.225	0.149	0.075	0.046
		Var 1	0.094	0.054	0.033	0.038
		Var 2	0.189	0.122	0.061	0.037
	.60	Var 1	0.086	0.086	0.058	0.067
		Var 2	0.123	0.111	0.073	0.058
Var = .5	.40	Default	0.183	0.152	0.124	0.126
		Var 1	0.122	0.088	0.052	0.045
		Var 2	0.162	0.130	0.096	0.086
	.60	Var 1	0.118	0.095	0.069	0.069
		Var 2	0.162	0.094	0.048	0.028
Test Length =40						
Var = 1	.40	Default	0.287	0.220	0.130	0.097
		Var 1	0.157	0.122	0.058	0.040
		Var 2	0.259	0.198	0.112	0.077
	.60	Var 1	0.073	0.099	0.055	0.045
		Var 2	0.121	0.186	0.128	0.101
Var = .5	.40	Default	0.257	0.250	0.240	0.204
		Var 1	0.152	0.137	0.086	0.060
		Var 2	0.229	0.210	0.172	0.148
	.60	Var 1	0.107	0.111	0.075	0.057
		Var 2	0.166	0.178	0.087	0.057

Table 4.4

Average Absolute Bias of *b*-parameter

a Prior	ρ	*b* Prior	Sample Size 100	200	500	1000
Test Length = 15						
Var = 1	.40	Default	0.296	0.305	0.310	0.320
		Var 1	0.277	0.289	0.292	0.310
		Var 2	0.287	0.296	0.301	0.313
	.60	Var 1	0.288	0.296	0.268	0.284
		Var 2	0.332	0.355	0.384	0.412
Var = .5	.40	Default	0.355	0.383	0.427	0.466
		Var 1	0.323	0.344	0.360	0.386
		Var 2	0.345	0.371	0.406	0.444
	.60	Var 1	0.288	0.194	0.327	0.344
		Var 2	0.276	0.285	0.289	0.301
Test Length = 25						
Var = 1	.40	Default	0.262	0.298	0.323	0.338
		Var 1	0.257	0.279	0.293	0.309
		Var 2	0.252	0.288	0.314	0.331
	.60	Var 1	0.280	0.308	0.313	0.328
		Var 2	0.254	0.337	0.398	0.433
Var = .5	.40	Default	0.316	0.370	0.444	0.494
		Var 1	0.291	0.326	0.366	0.391
		Var 2	0.305	0.354	0.421	0.464
	.60	Var 1	0.280	0.340	0.368	0.391
		Var 2	0.211	0.274	0.301	0.318
Test Length =40						
Var = 1	.40	Default	0.310	0.314	0.336	0.351
		Var 1	0.304	0.301	0.306	0.312
		Var 2	0.300	0.304	0.327	0.339
	.60	Var 1	0.309	0.319	0.319	0.326
		Var 2	0.281	0.362	0.400	0.430
Var = .5	.40	Default	0.366	0.399	0.464	0.501
		Var 1	0.336	0.346	0.375	0.393
		Var 2	0.353	0.377	0.431	0.470
	.60	Var 1	0.301	0.339	0.372	0.386
		Var 2	0.234	0.298	0.310	0.322

Table 4.5

Average Standard Deviation of *a*-parameter

a Prior	ρ	*b* Prior	Sample Size 100	200	500	1000
Test Length = 15						
Var = 1	.40	Default	0.334	0.241	0.151	0.101
		Var 1	0.282	0.221	0.138	0.095
		Var 2	0.320	0.240	0.151	0.101
	.60	Var 1	0.220	0.188	0.148	0.095
		Var 2	0.181	0.162	0.124	0.095
Var = .5	.40	Default	0.189	0.169	0.138	0.114
		Var 1	0.165	0.151	0.113	0.087
		Var 2	0.185	0.166	0.130	0.101
	.60	Var 1	0.252	0.156	0.111	0.085
		Var 2	0.325	0.239	0.150	0.100
Test Length = 25						
Var = 1	.40	Default	0.320	0.227	0.124	0.085
		Var 1	0.248	0.179	0.106	0.073
		Var 2	0.297	0.215	0.121	0.082
	.60	Var 1	0.321	0.180	0.102	0.068
		Var 2	0.182	0.152	0.104	0.081
Var = .5	.40	Default	0.182	0.166	0.123	0.107
		Var 1	0.153	0.131	0.092	0.069
		Var 2	0.176	0.154	0.111	0.090
	.60	Var 1	0.159	0.125	0.086	0.064
		Var 2	0.323	0.209	0.118	0.080
Test Length =40						
Var = 1	.40	Default	0.342	0.252	0.132	0.091
		Var 1	0.260	0.198	0.109	0.077
		Var 2	0.316	0.240	0.127	0.087
	.60	Var 1	0.245	0.208	0.104	0.073
		Var 2	0.176	0.163	0.112	0.085
Var = .5	.40	Default	0.194	0.185	0.148	0.109
		Var 1	0.161	0.146	0.100	0.075
		Var 2	0.184	0.170	0.125	0.095
	.60	Var 1	0.154	0.137	0.094	0.070
		Var 2	0.312	0.231	0.121	0.083

Table 4.6

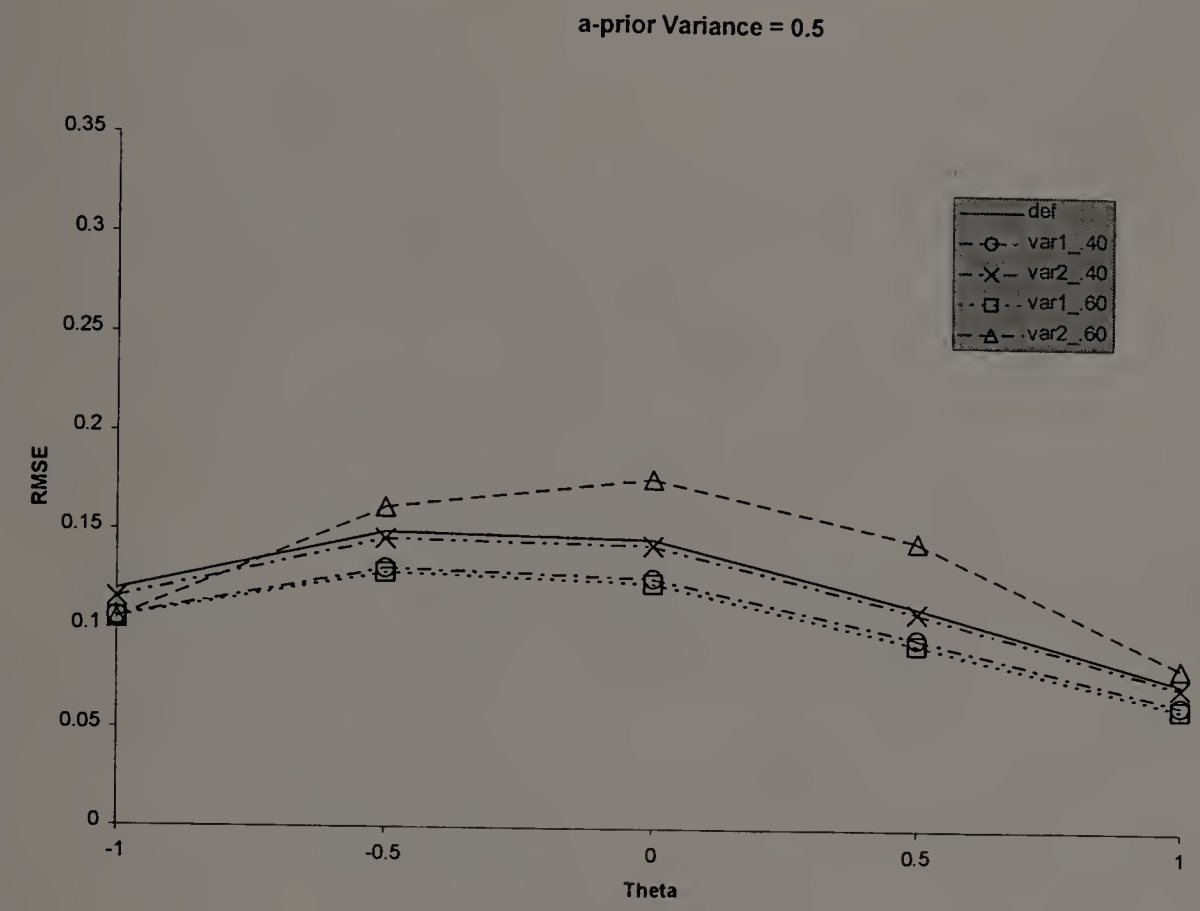
Average Standard Deviation of *b*-parameter

a Prior	ρ	*b* Prior	Sample Size 100	200	500	1000
Test Length = 15						
Var = 1	.40	Default	0.321	0.254	0.193	0.148
		Var 1	0.271	0.215	0.158	0.121
		Var 2	0.311	0.246	0.184	0.143
	.60	Var 1	0.258	0.203	0.147	0.114
		Var 2	0.282	0.226	0.178	0.141
Var = .5	.40	Default	0.299	0.242	0.202	0.170
		Var 1	0.263	0.209	0.159	0.125
		Var 2	0.291	0.233	0.189	0.154
	.60	Var 1	0.258	0.191	0.149	0.118
		Var 2	0.300	0.238	0.176	0.135
Test Length = 25						
Var = 1	.40	Default	0.342	0.271	0.201	0.158
		Var 1	0.295	0.229	0.168	0.133
		Var 2	0.332	0.260	0.192	0.151
	.60	Var 1	0.307	0.224	0.164	0.133
		Var 2	0.288	0.243	0.186	0.154
Var = .5	.40	Default	0.327	0.262	0.209	0.183
		Var 1	0.292	0.226	0.171	0.138
		Var 2	0.319	0.251	0.195	0.164
	.60	Var 1	0.264	0.225	0.168	0.138
		Var 2	0.301	0.250	0.185	0.145
Test Length =40						
Var = 1	.40	Default	0.339	0.277	0.198	0.154
		Var 1	0.283	0.230	0.165	0.134
		Var 2	0.327	0.266	0.193	0.151
	.60	Var 1	0.314	0.233	0.157	0.129
		Var 2	0.288	0.245	0.183	0.151
Var = .5	.40	Default	0.319	0.267	0.202	0.164
		Var 1	0.279	0.226	0.166	0.139
		Var 2	0.309	0.254	0.194	0.159
	.60	Var 1	0.269	0.213	0.158	0.133
		Var 2	0.297	0.257	0.183	0.145

Figure 4.1

RMSE Between True and Estimated Information Functions

N=100, n=15



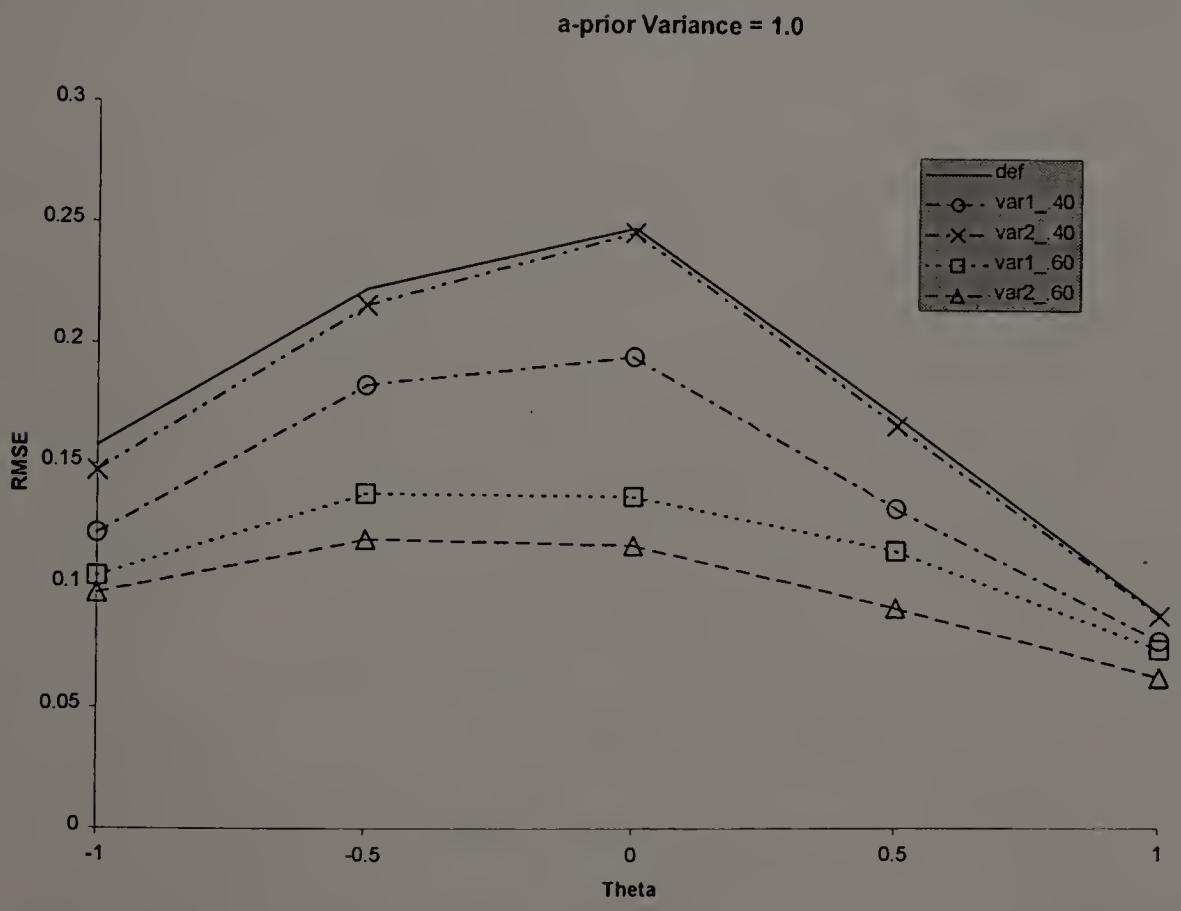
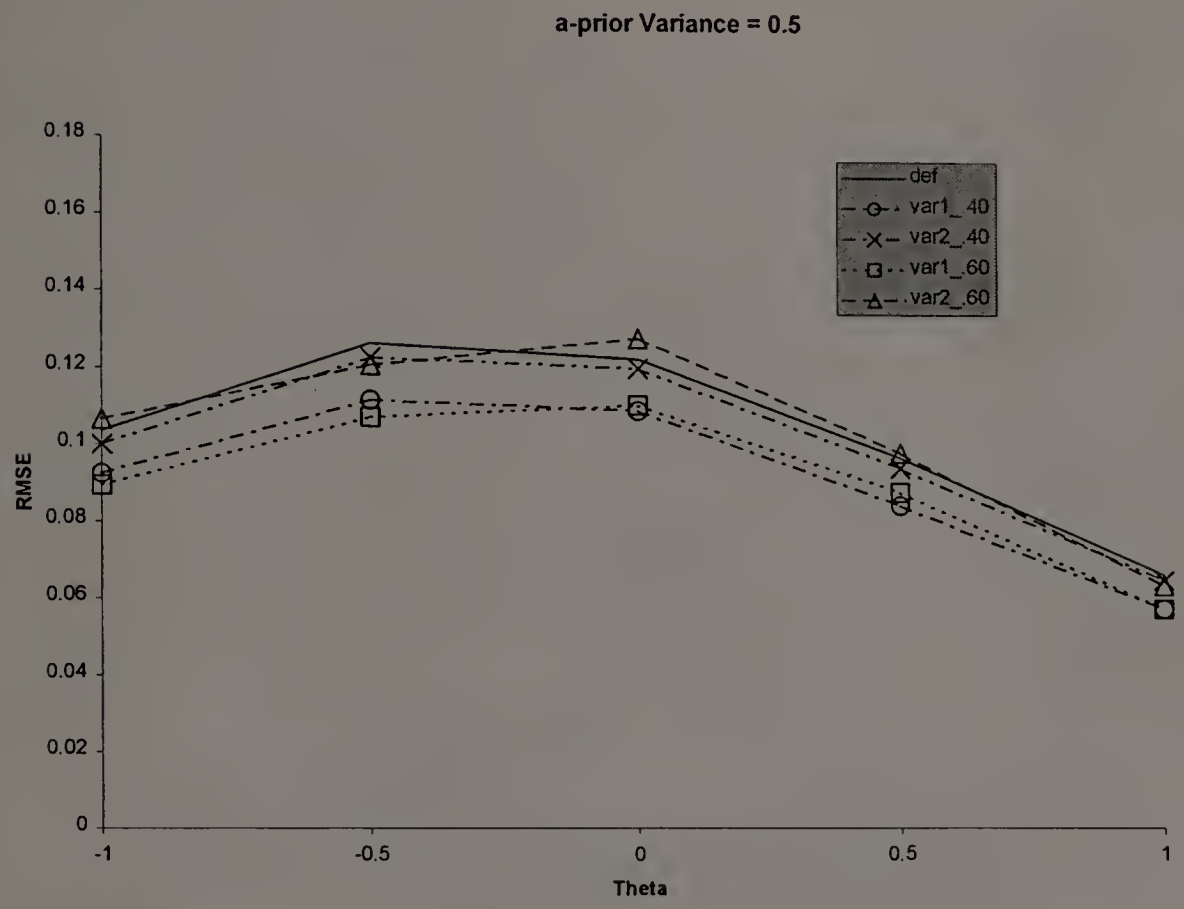


Figure 4.2

RMSE Between True and Estimated Information Functions

N=200, n=15



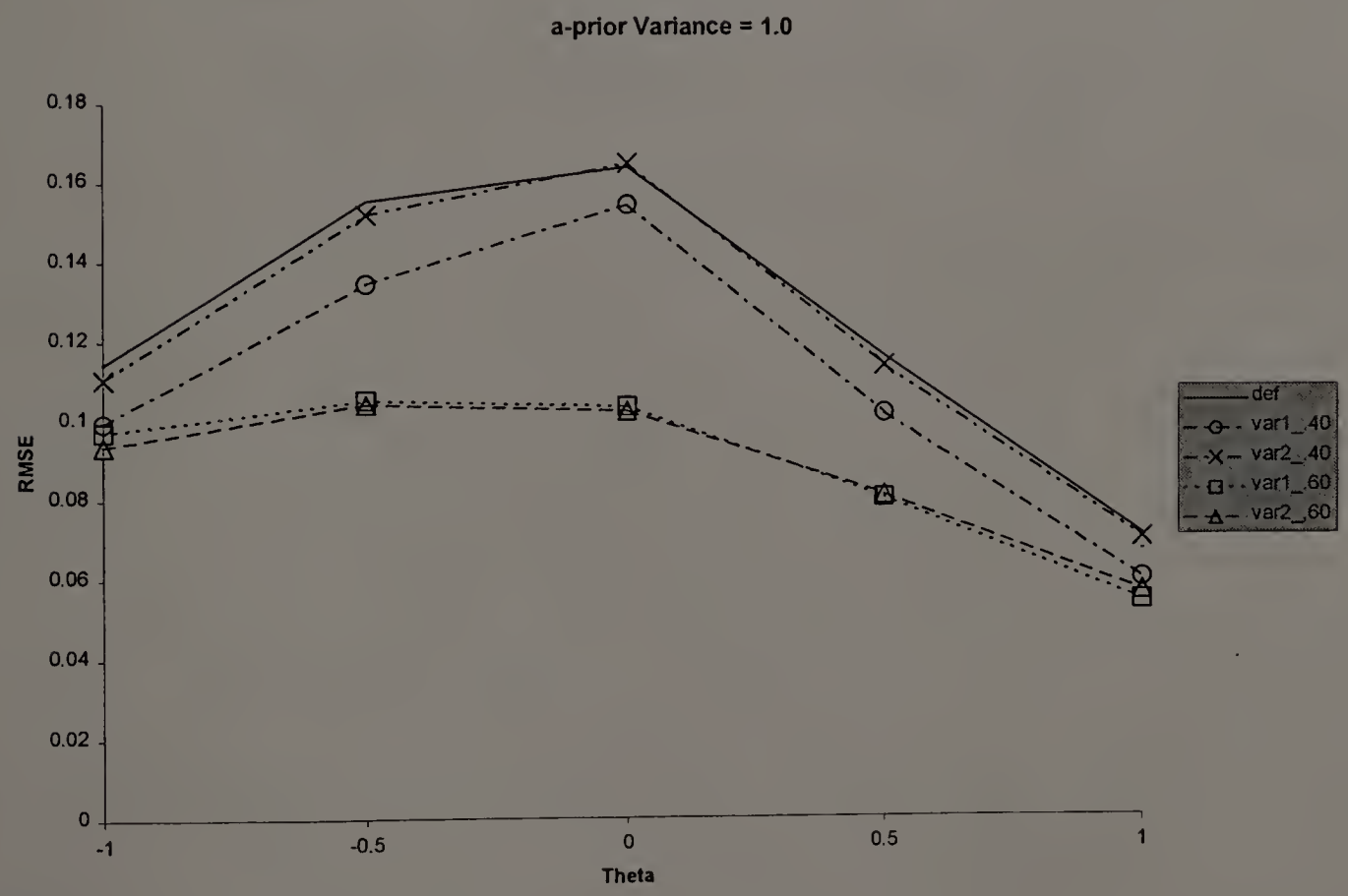
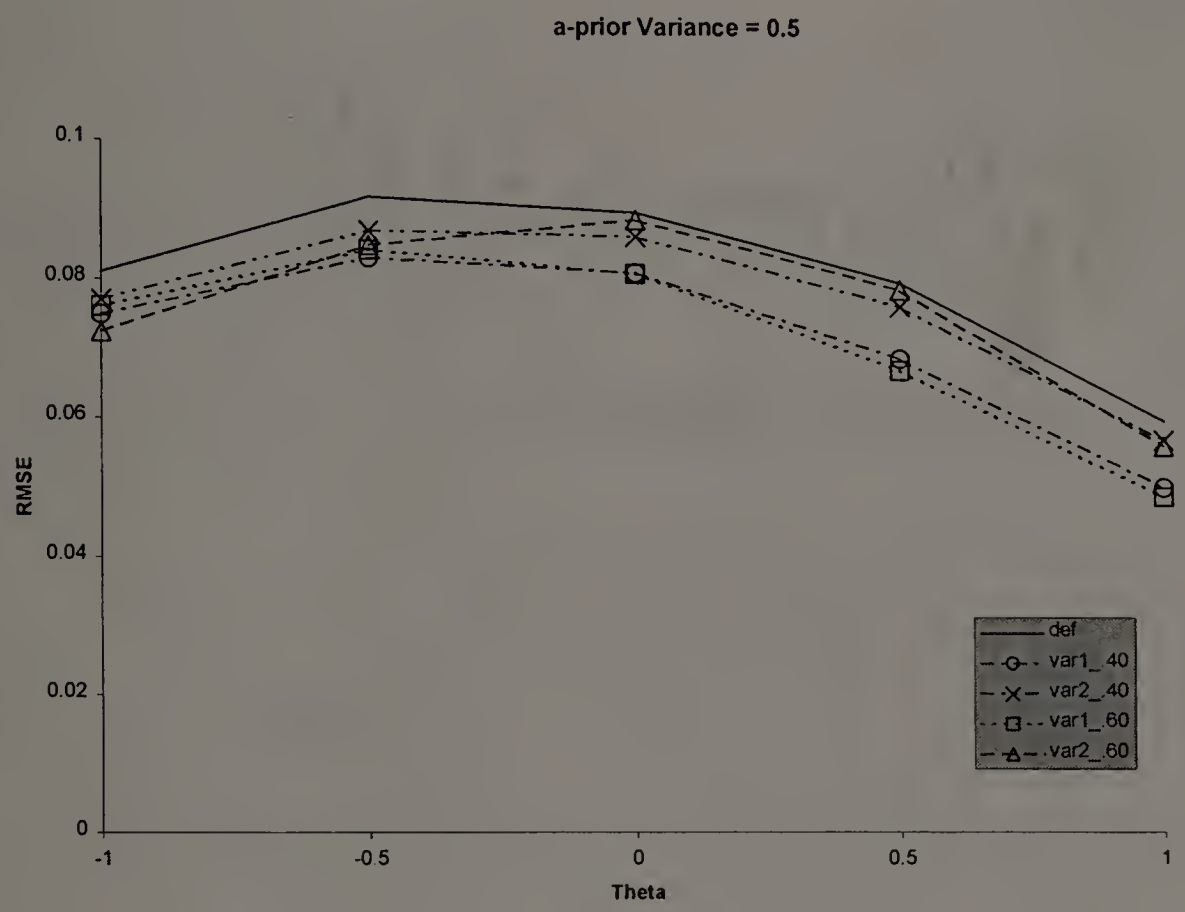


Figure 4.3

RMSE Between True and Estimated Information Functions

N=500, n=15



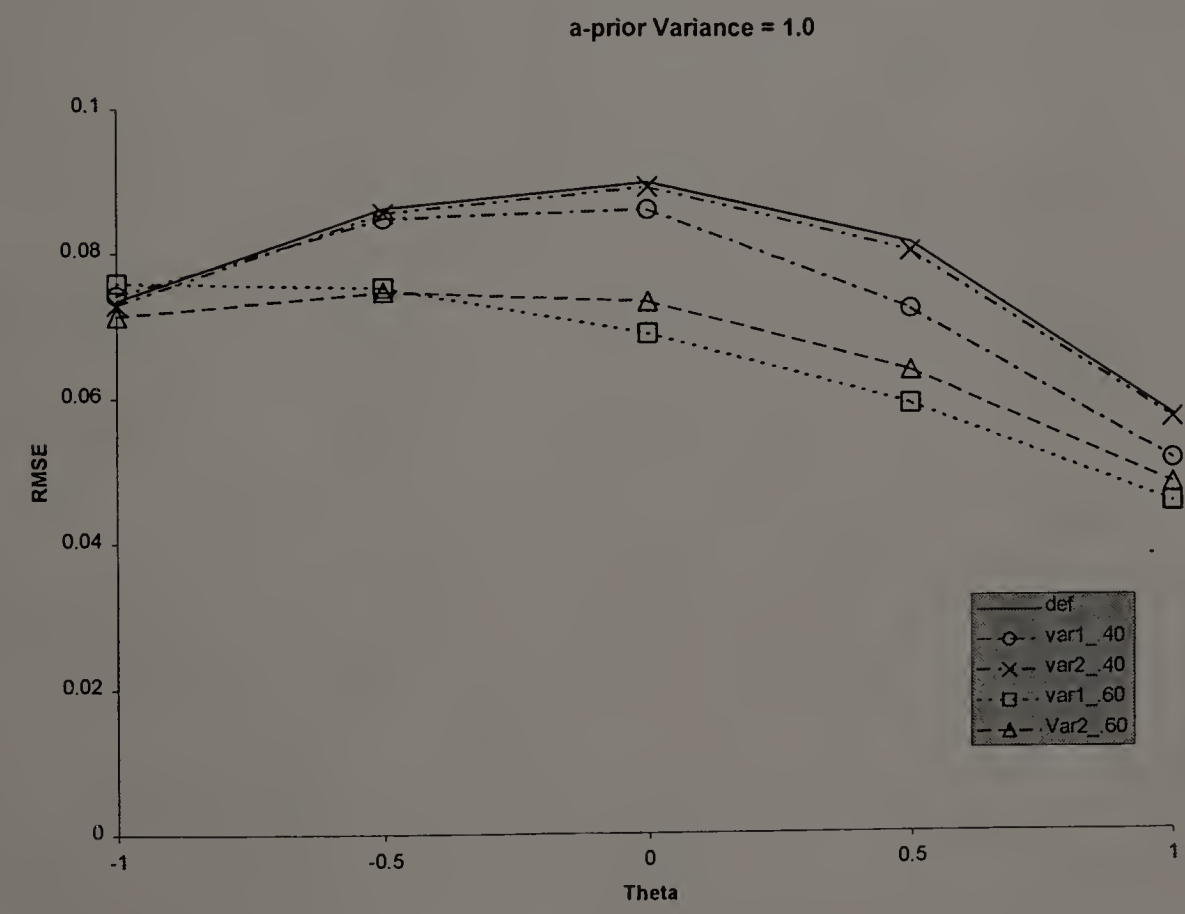
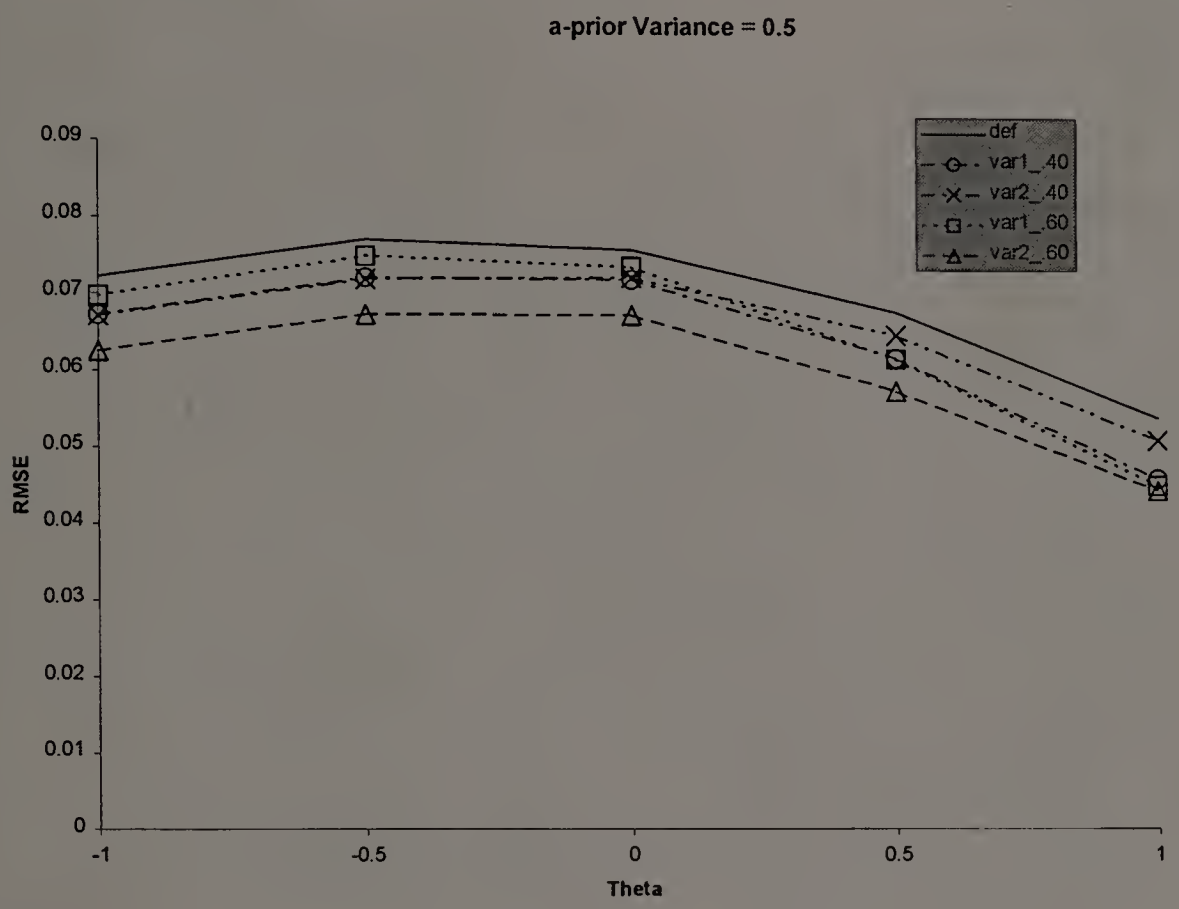


Figure 4.4

RMSE Between True and Estimated Information Functions

N=1000, n=15



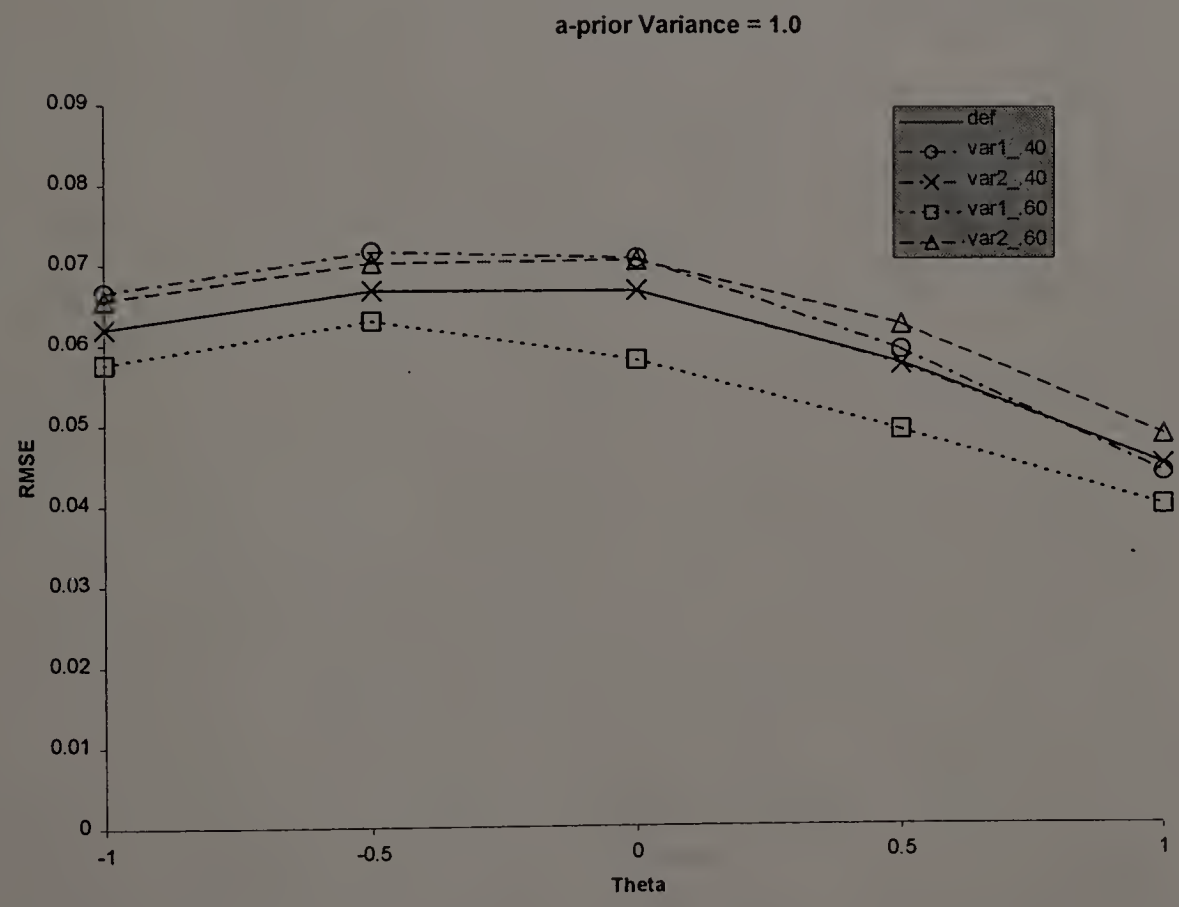
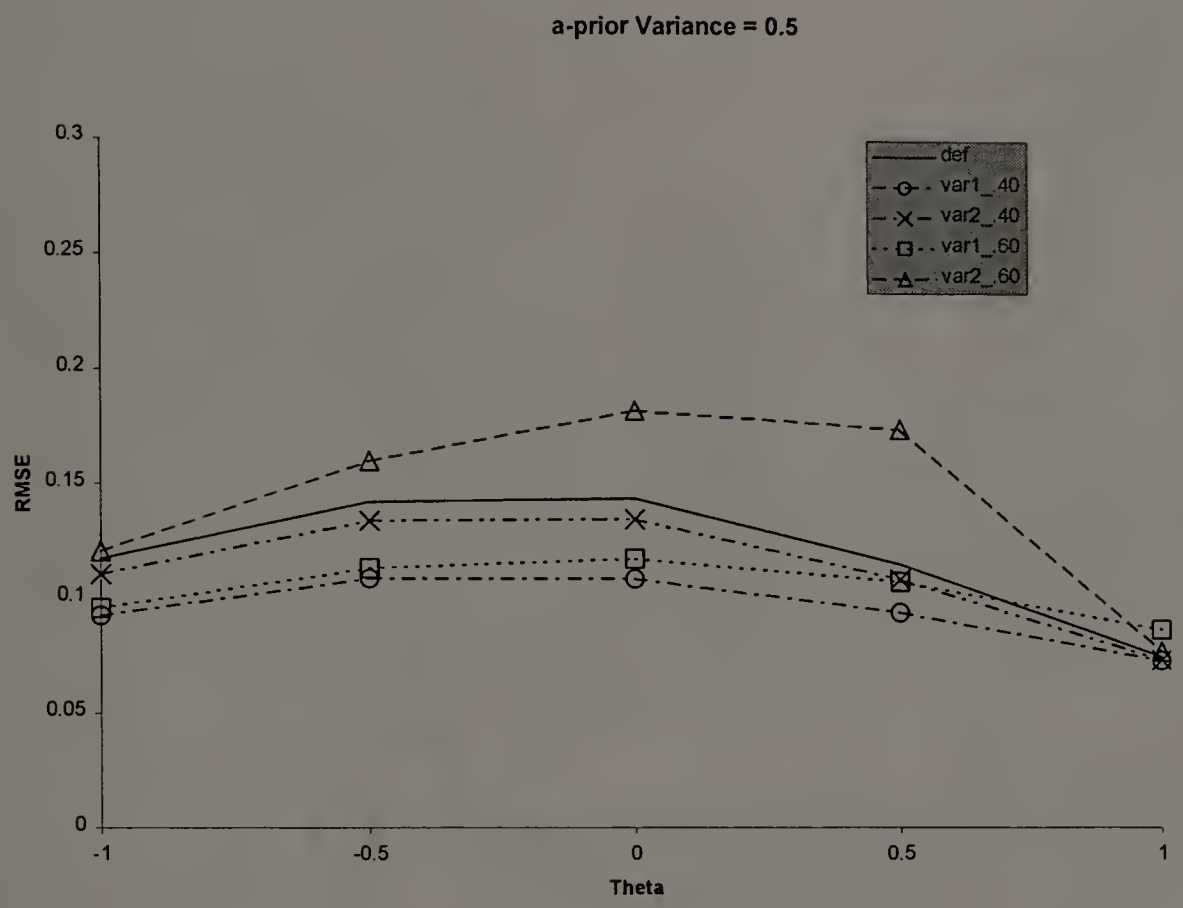


Figure 4.5

RMSE Between True and Estimated Information Functions

N=100, n=25



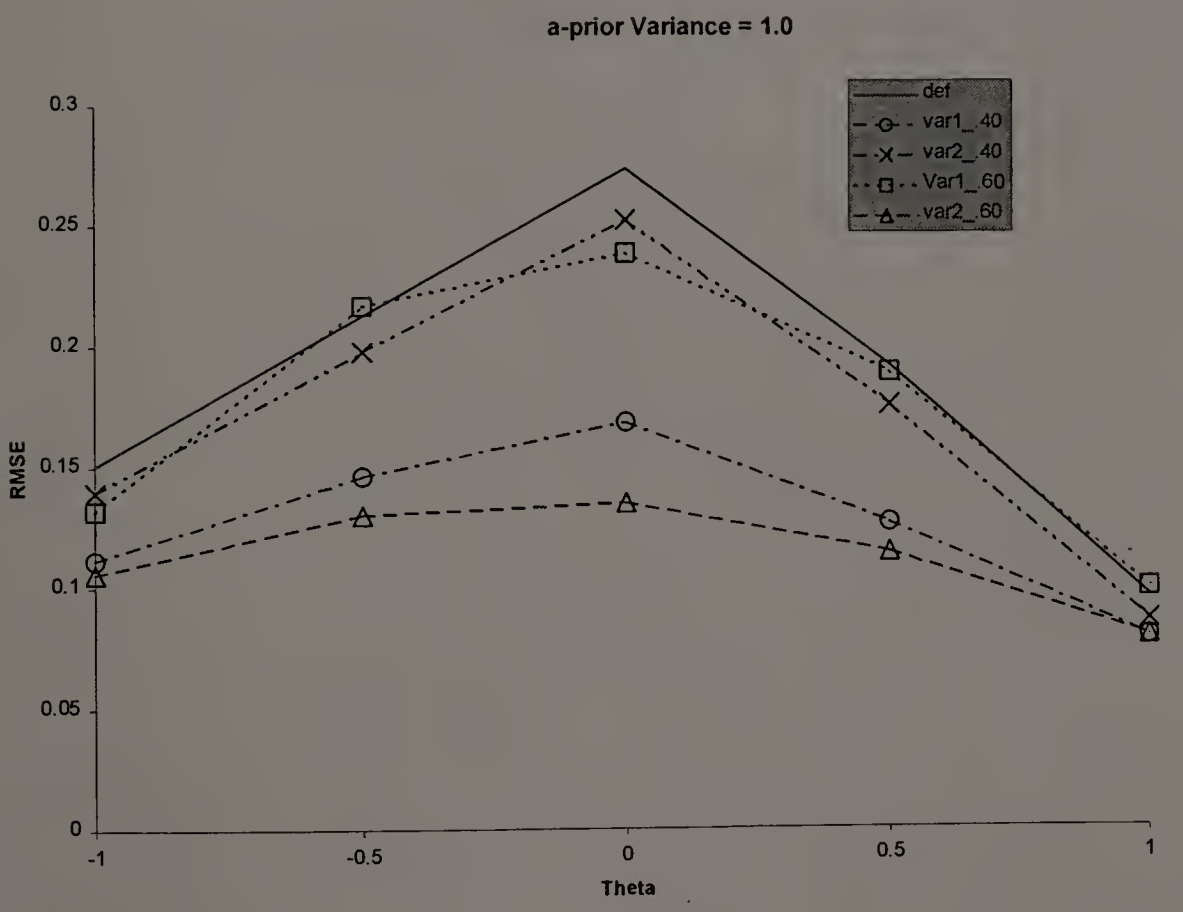
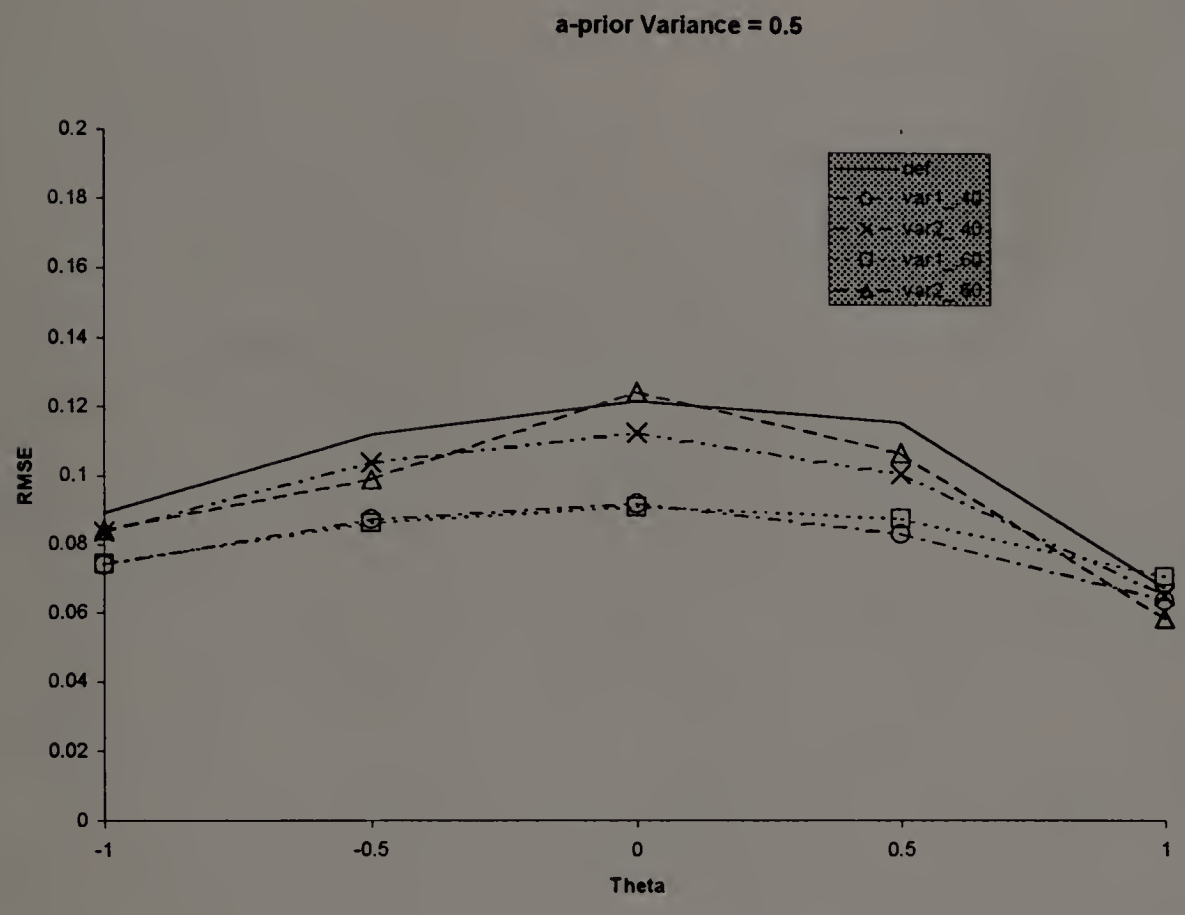


Figure 4.6

RMSE Between True and Estimated Information Functions

N=200, n=25



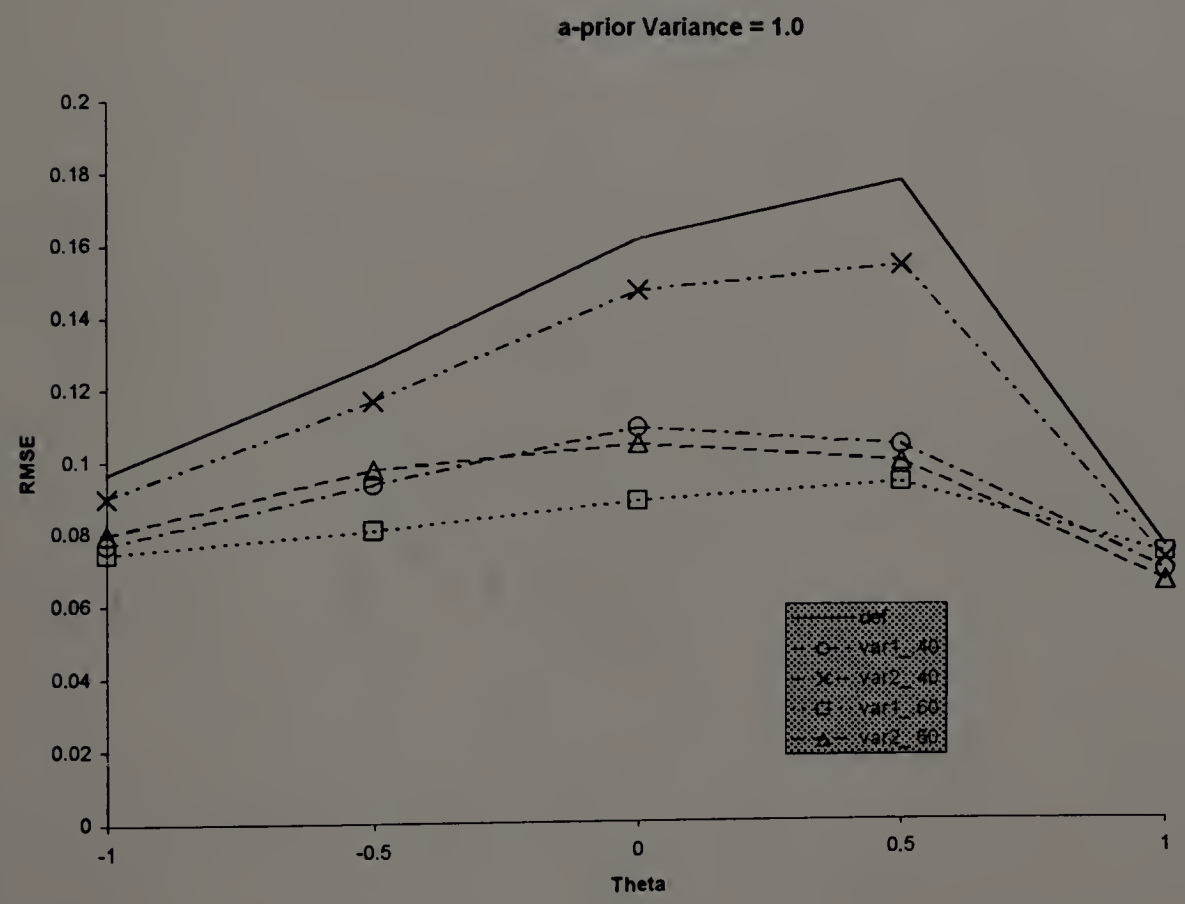
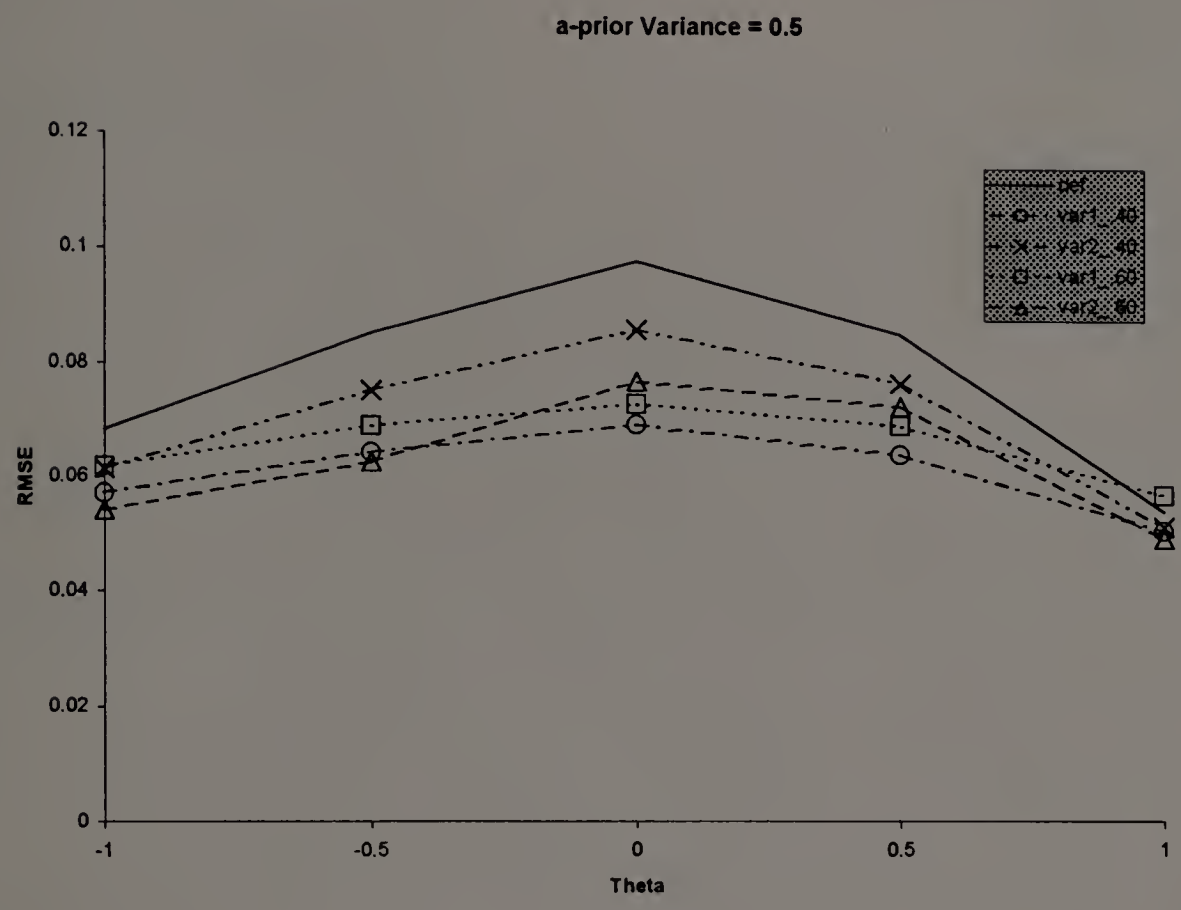


Figure 4.7

RMSE Between True and Estimated Information Functions

N=500, n=25



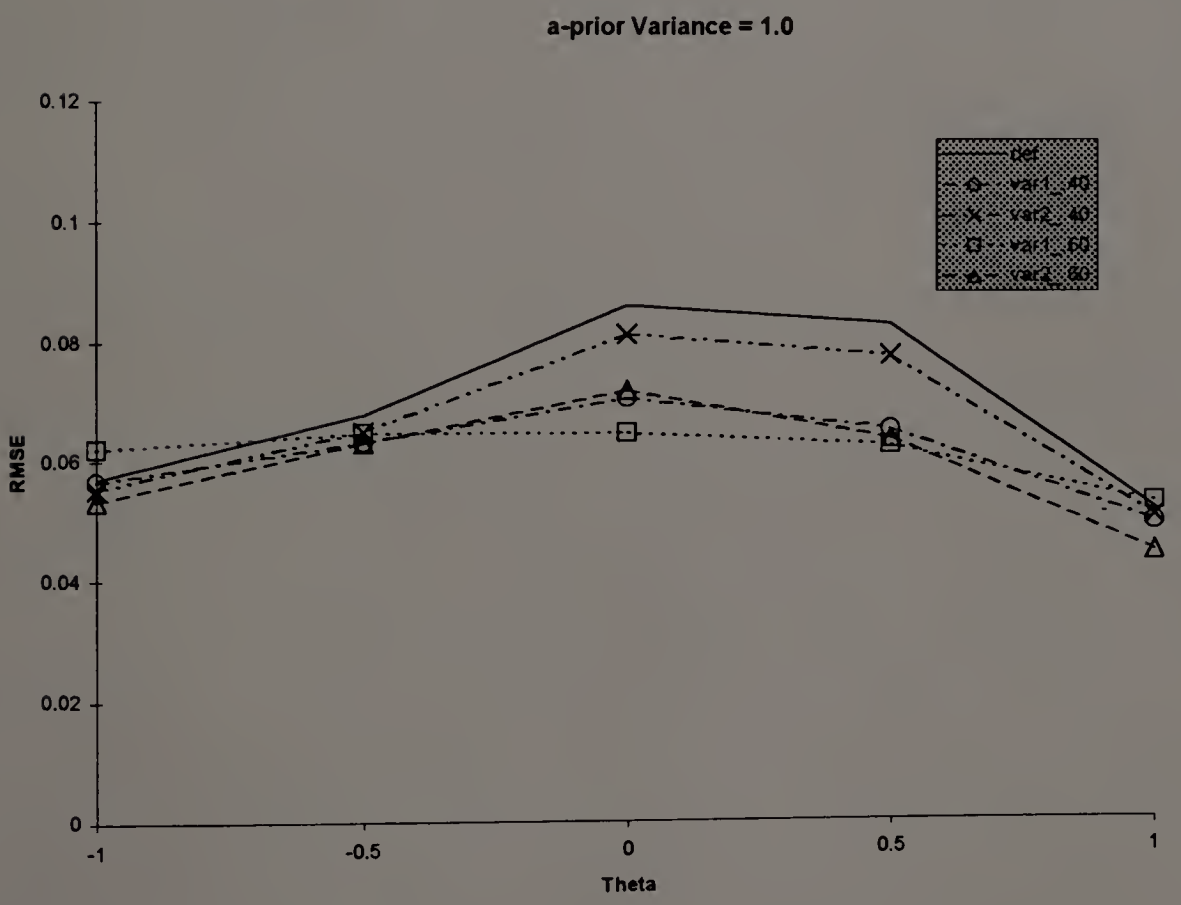
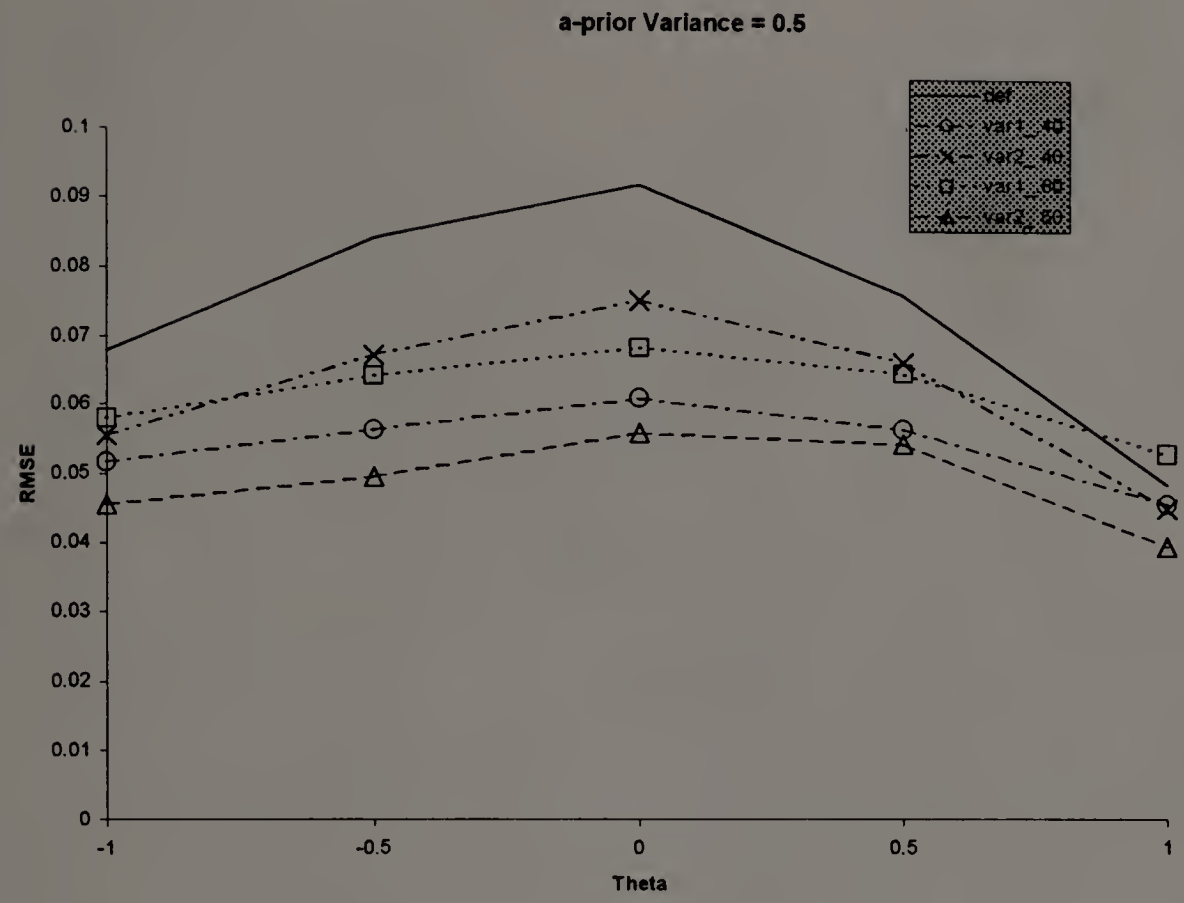


Figure 4.8

RMSE Between True and Estimated Information Functions

N=1000, n=25



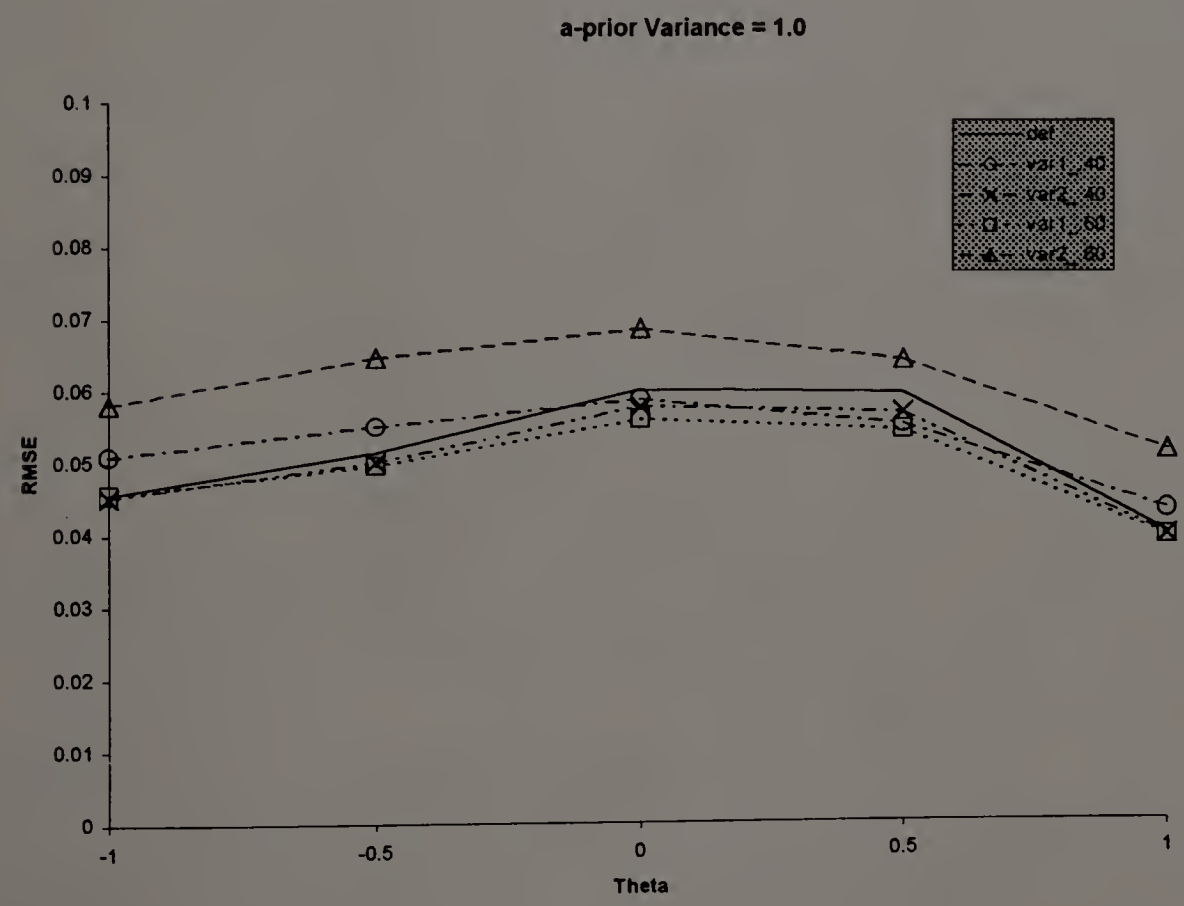
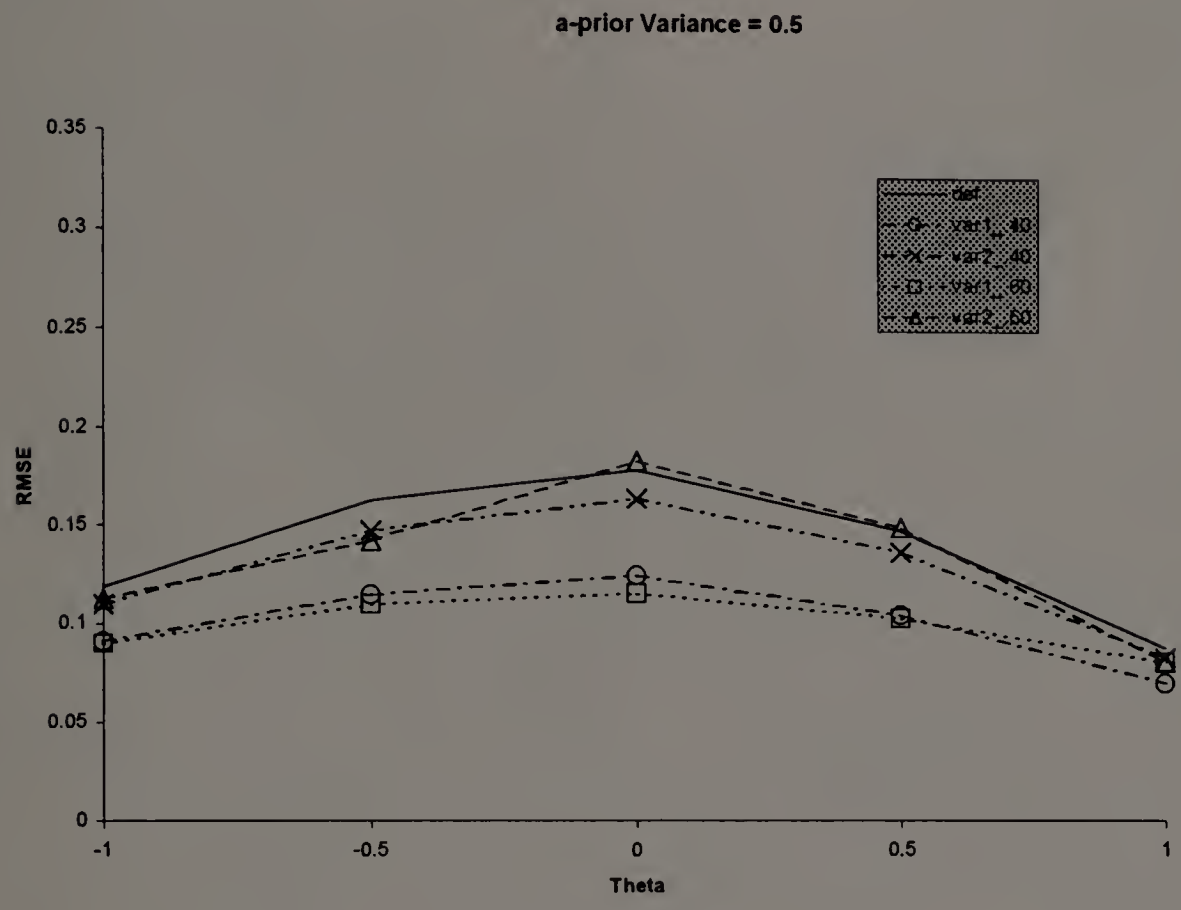


Figure 4.9

RMSE Between True and Estimated Information Functions

N=100, n=40



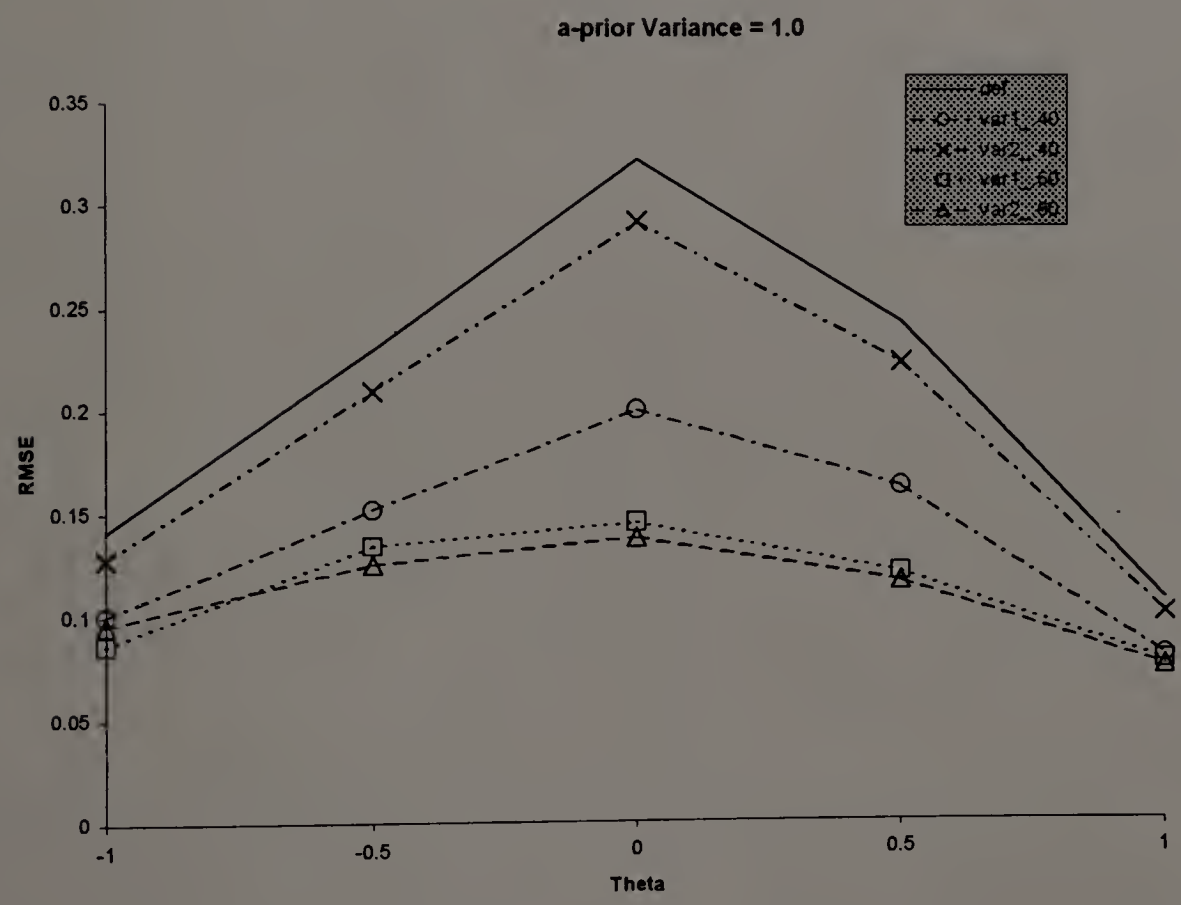
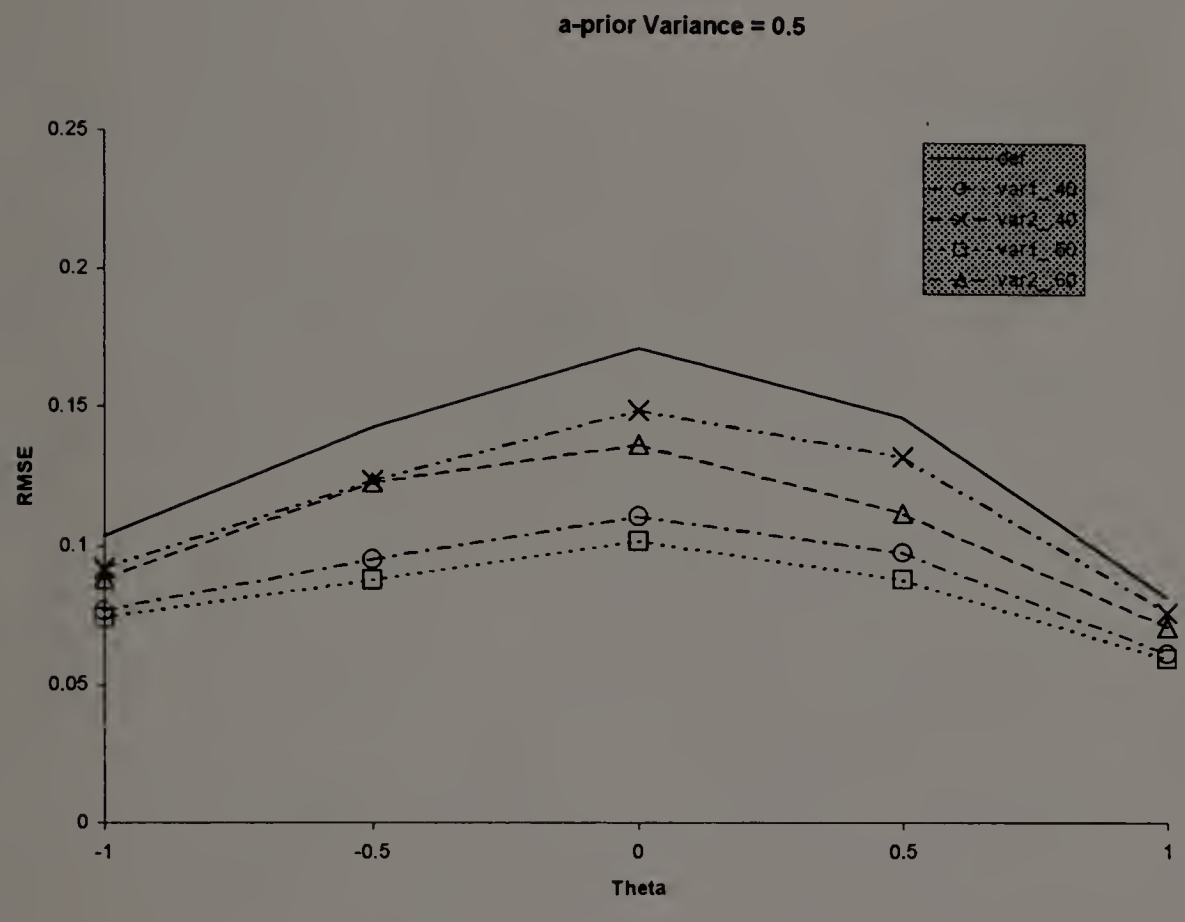


Figure 4.10

RMSE Between True and Estimated Information Functions

N=200, n=40



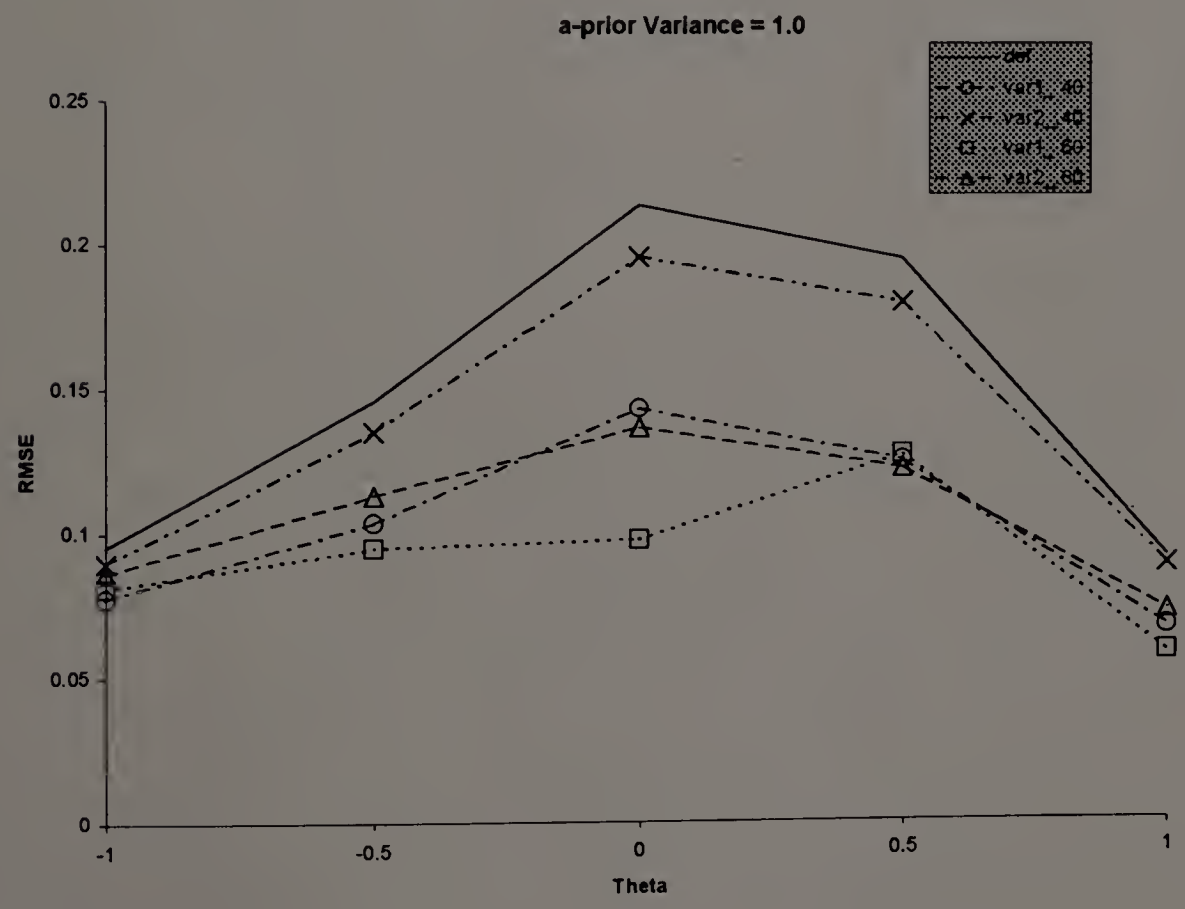
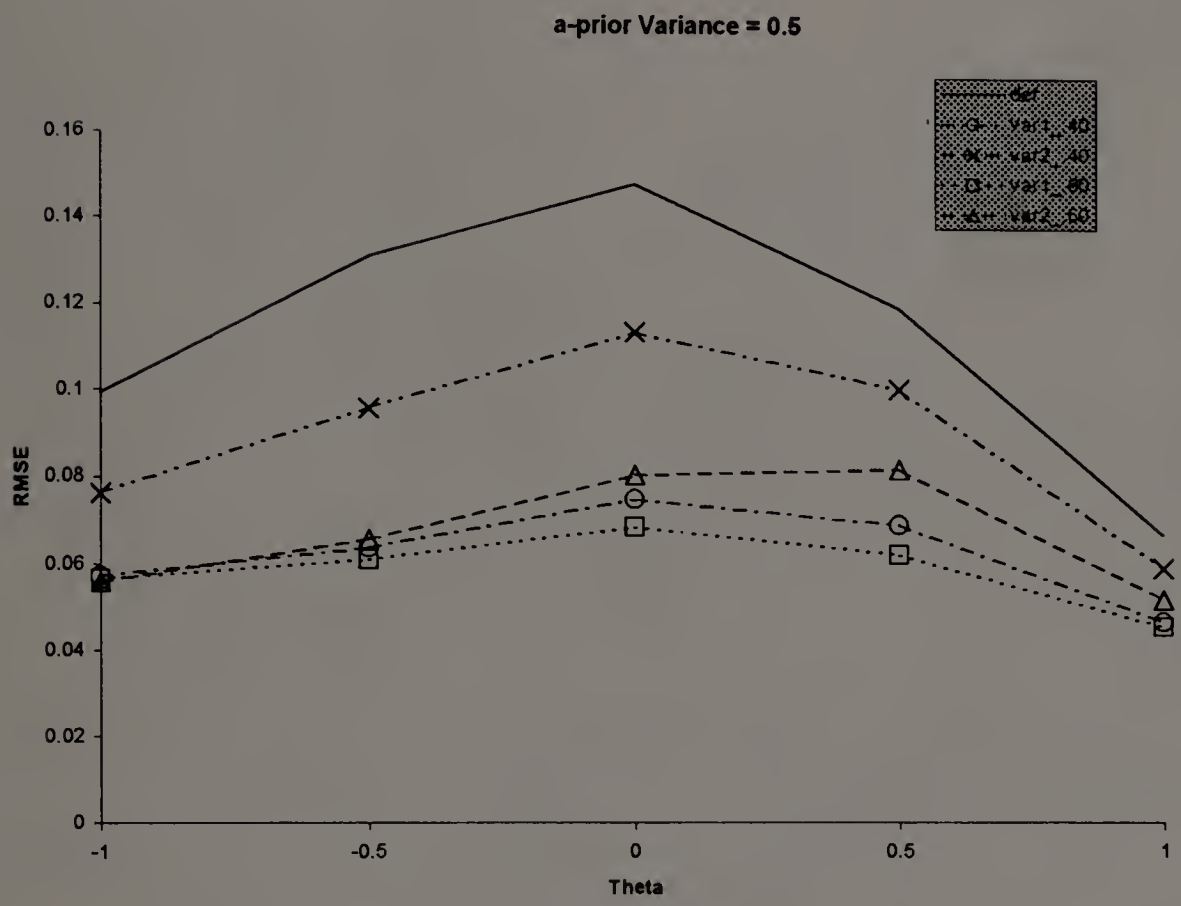


Figure 4.11

RMSE Between True and Estimated Information Functions

N=500, n=40



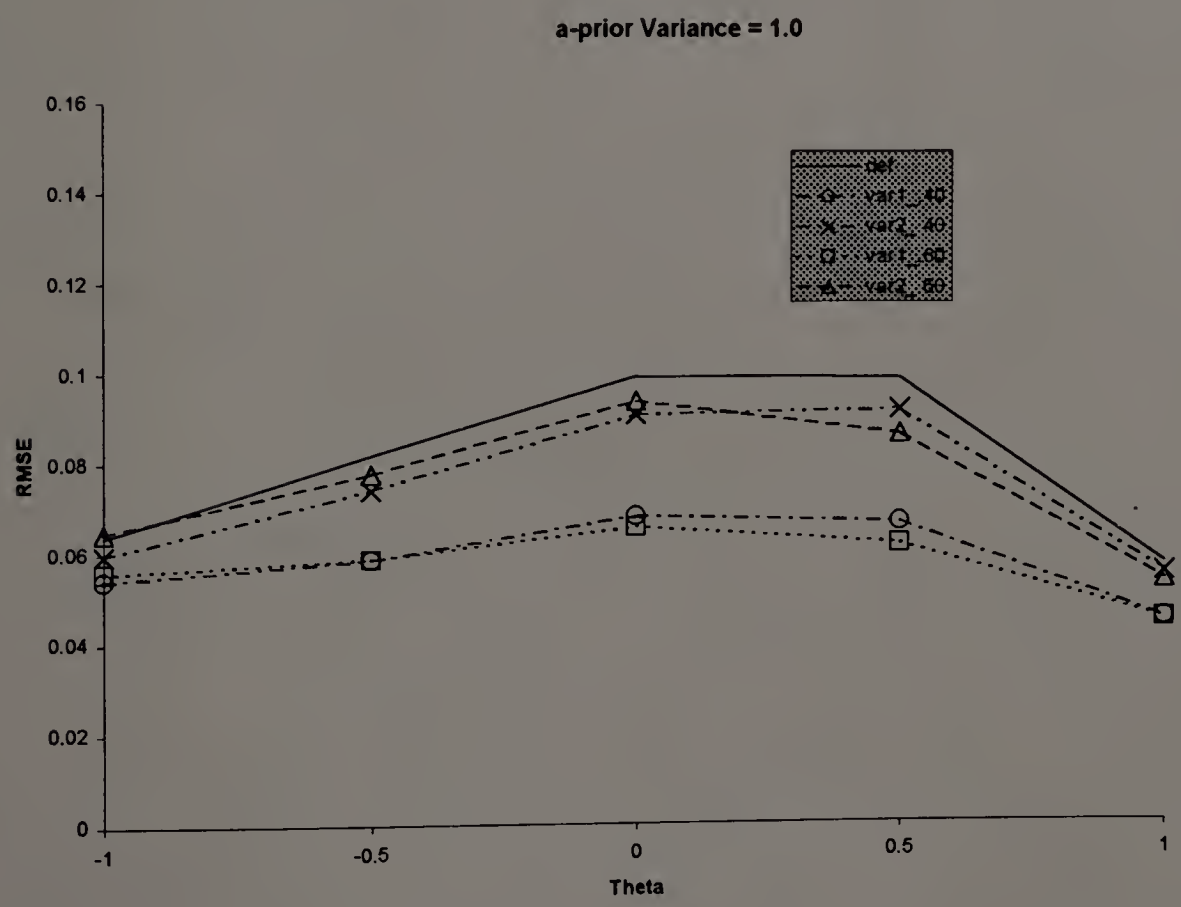
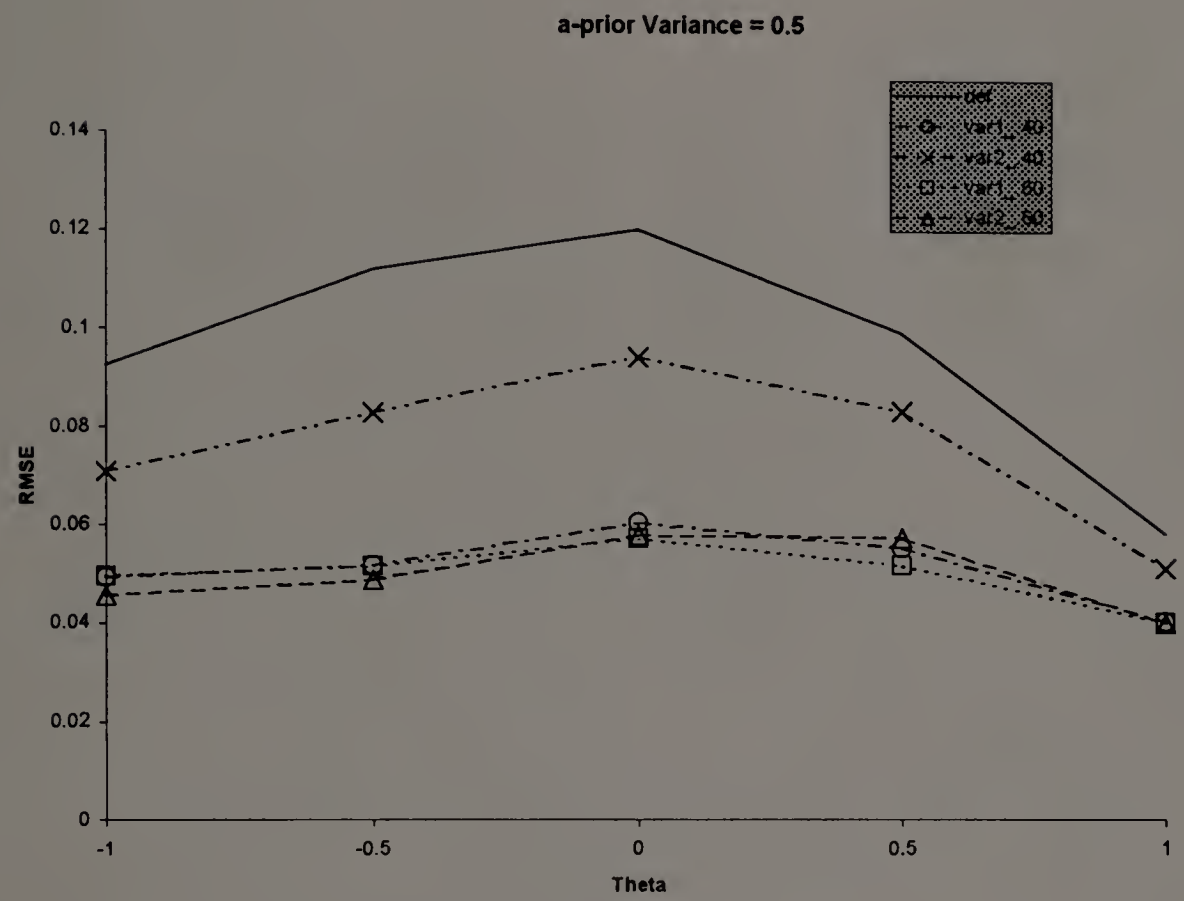


Figure 4.12

RMSE Between True and Estimated Information Functions

N=1000, n=40



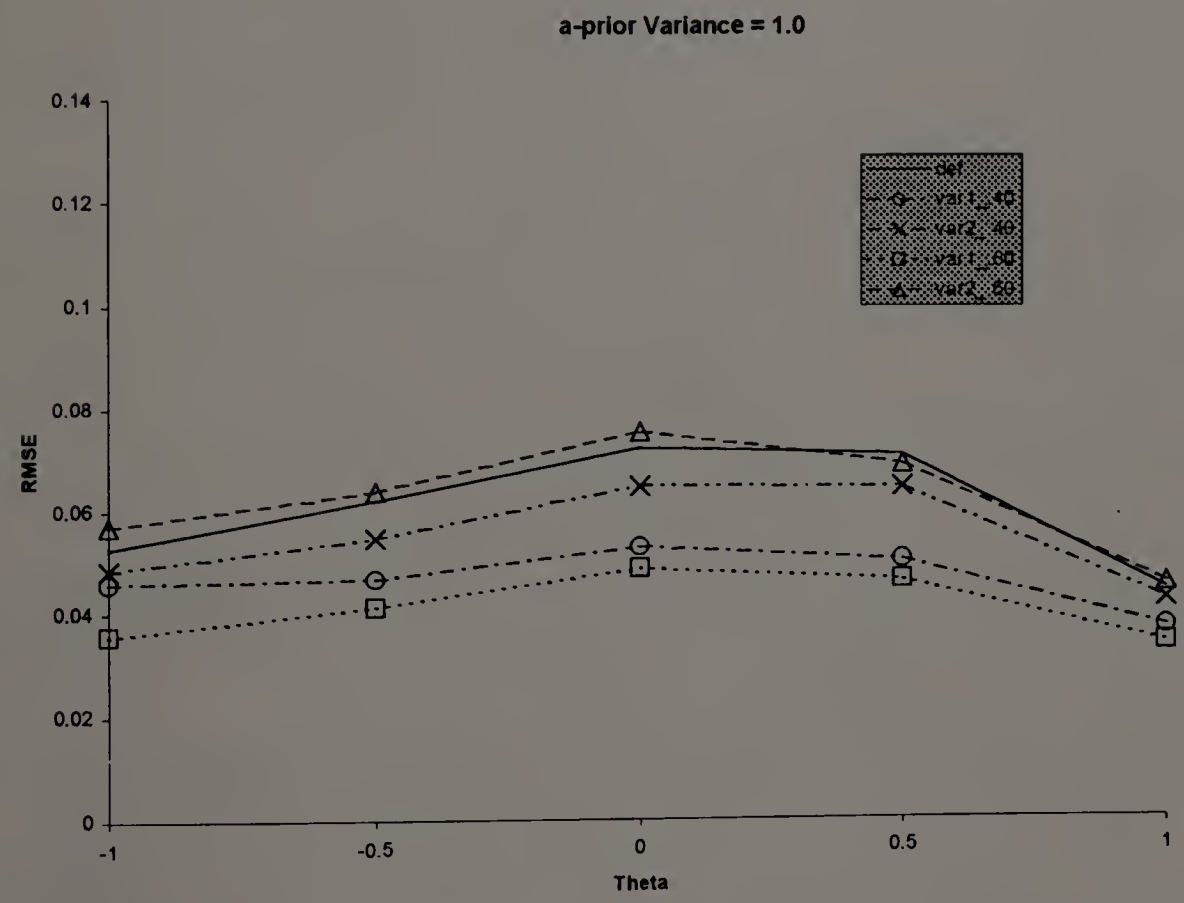
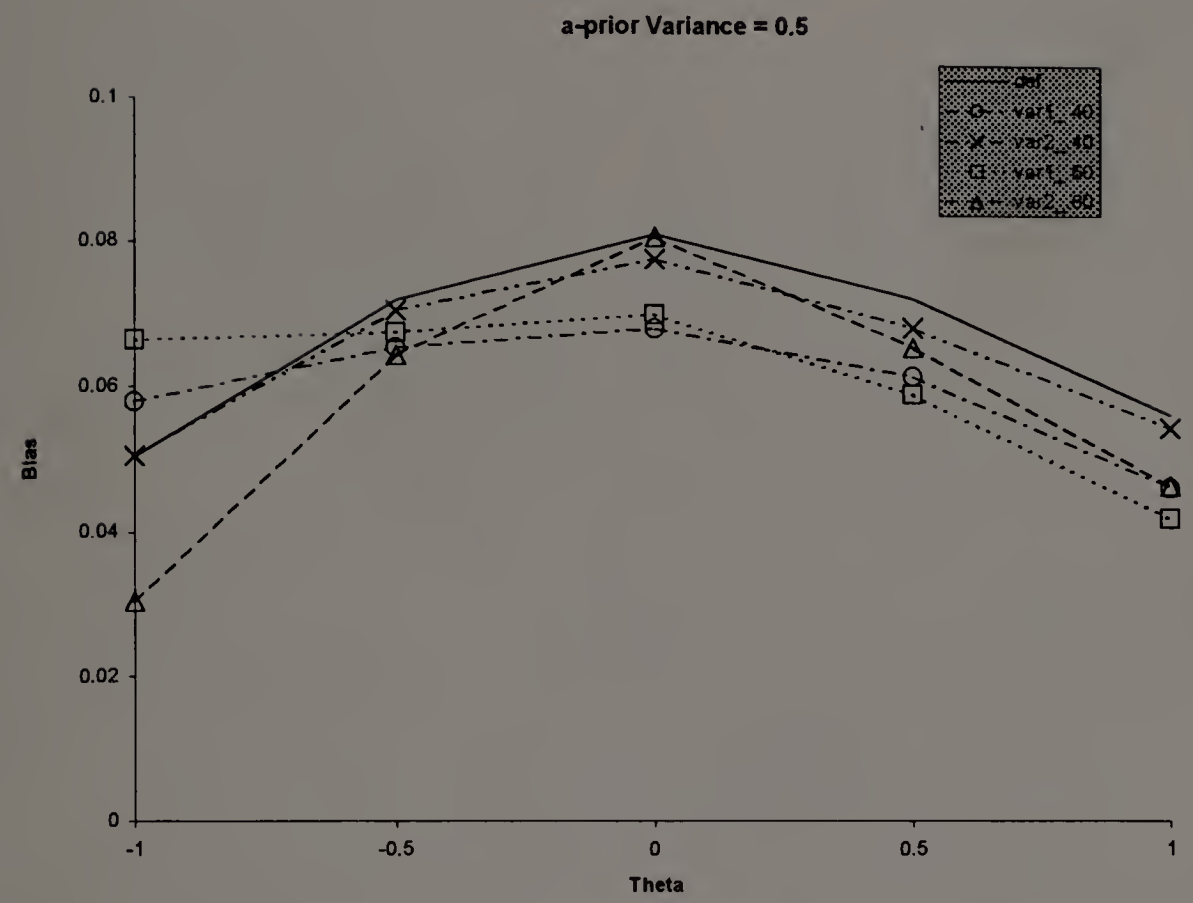


Figure 4.13

Bias of Estimated Information Functions

N=100, n=15



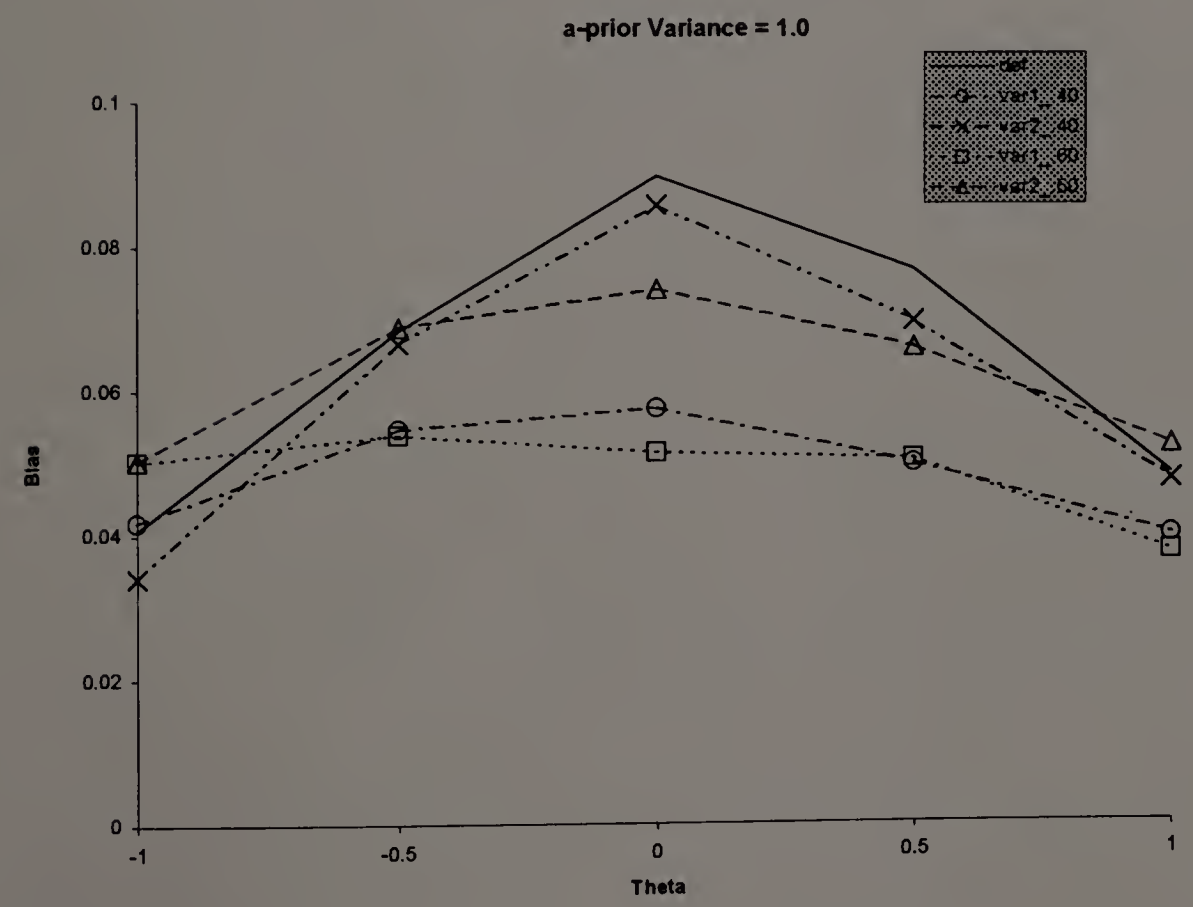
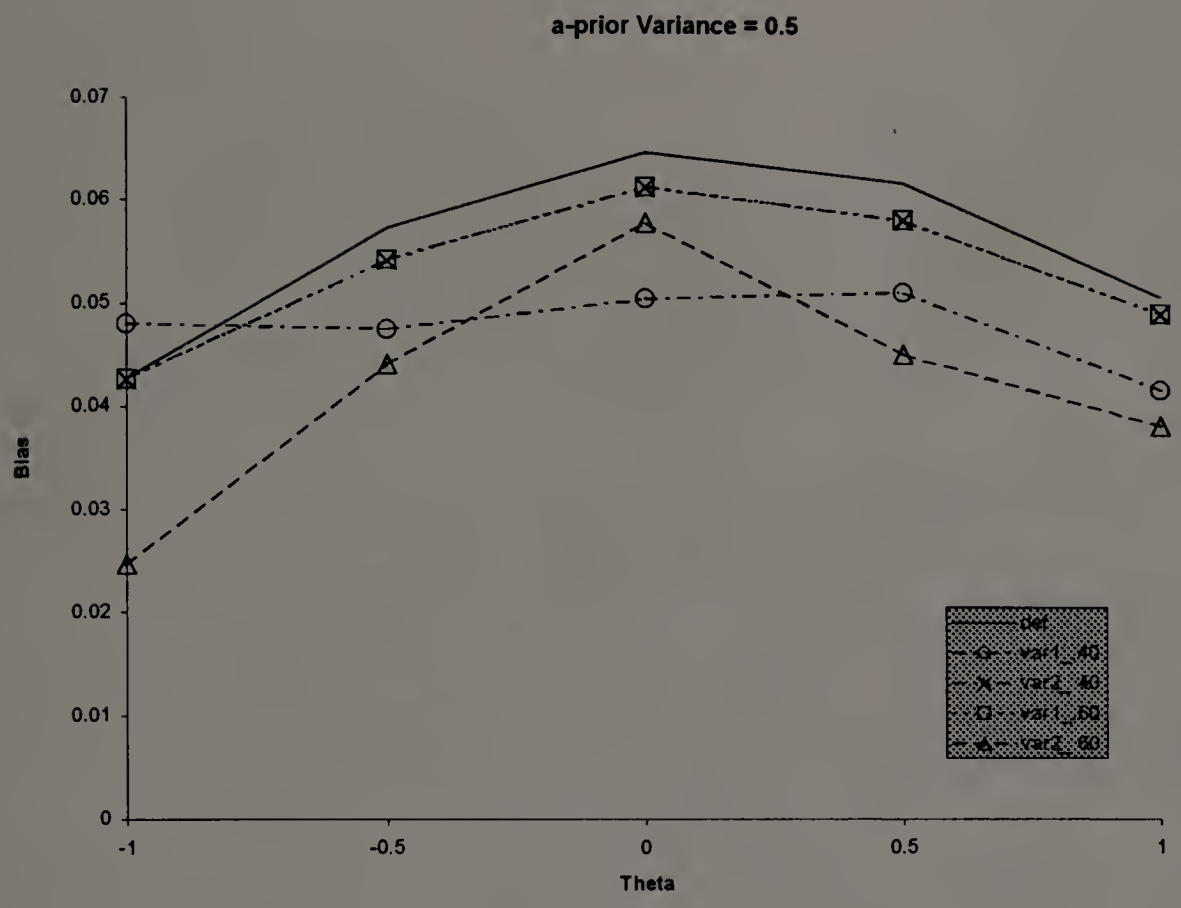


Figure 4.14

Bias of Estimated Information Functions

N=200, n=15



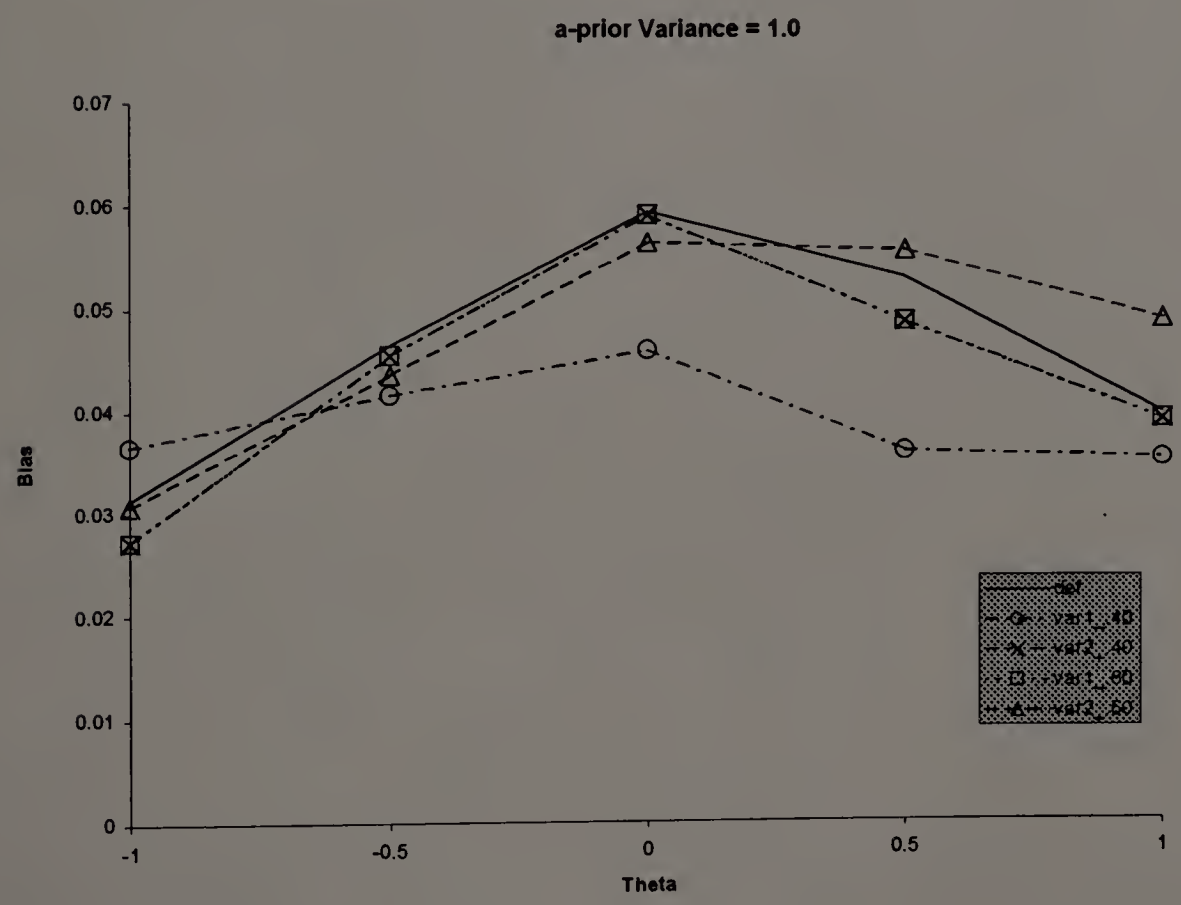
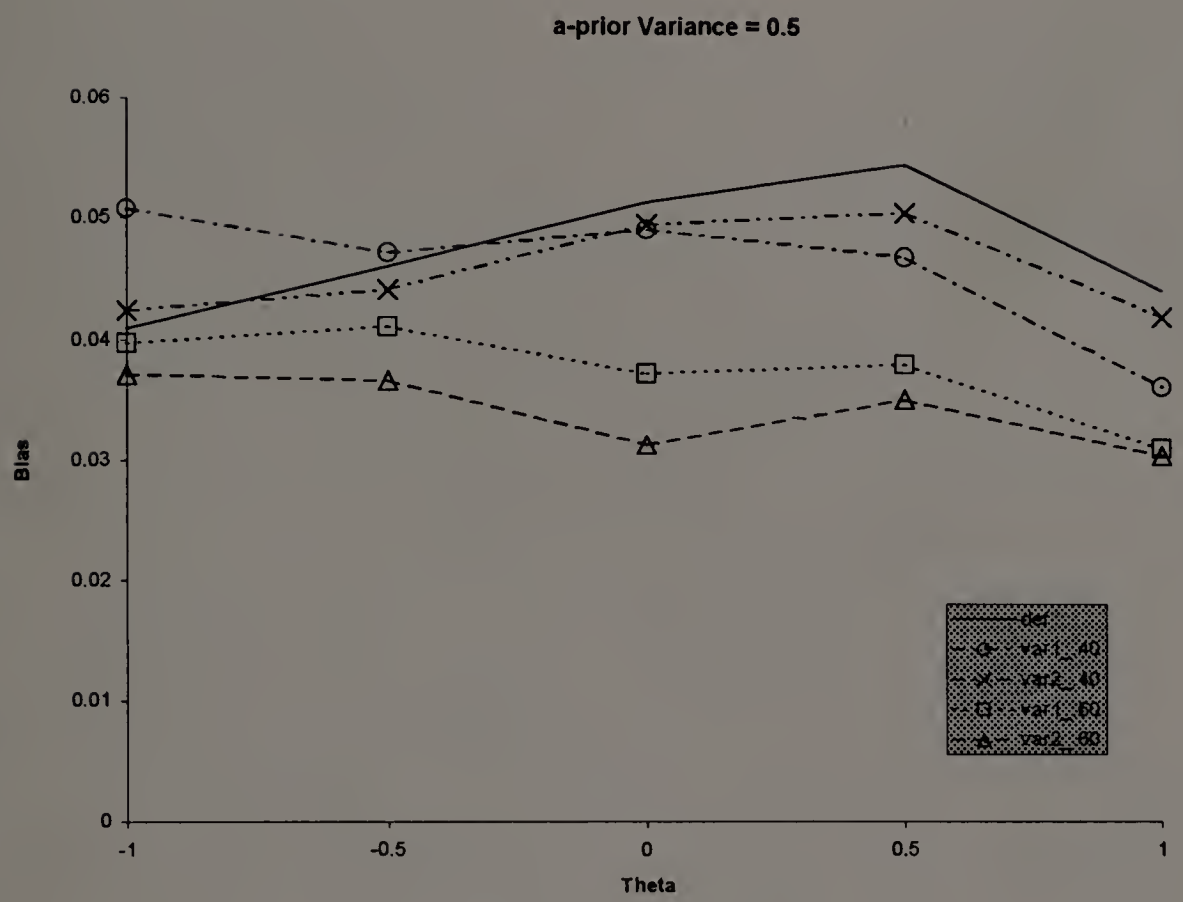


Figure 4.15

Bias of Estimated Information Functions

N=500, n=15



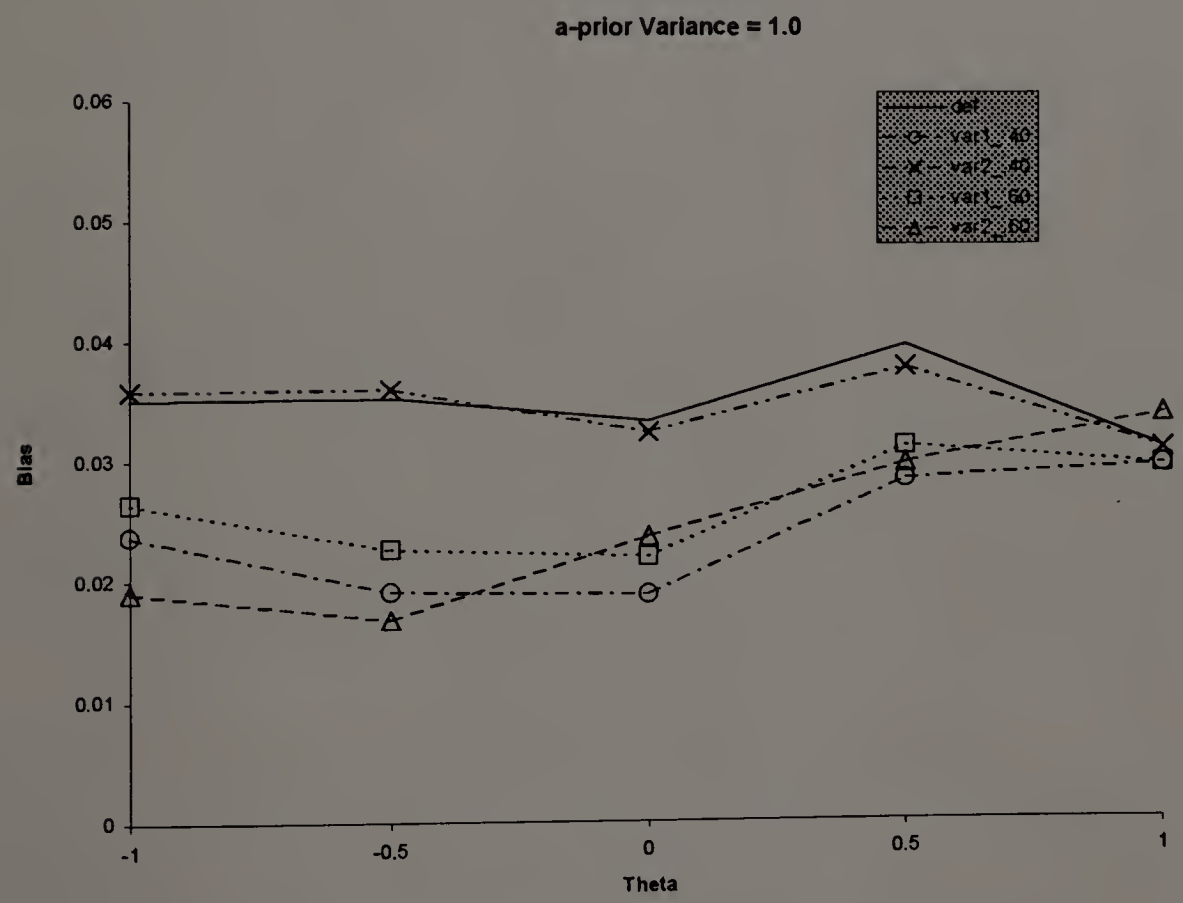
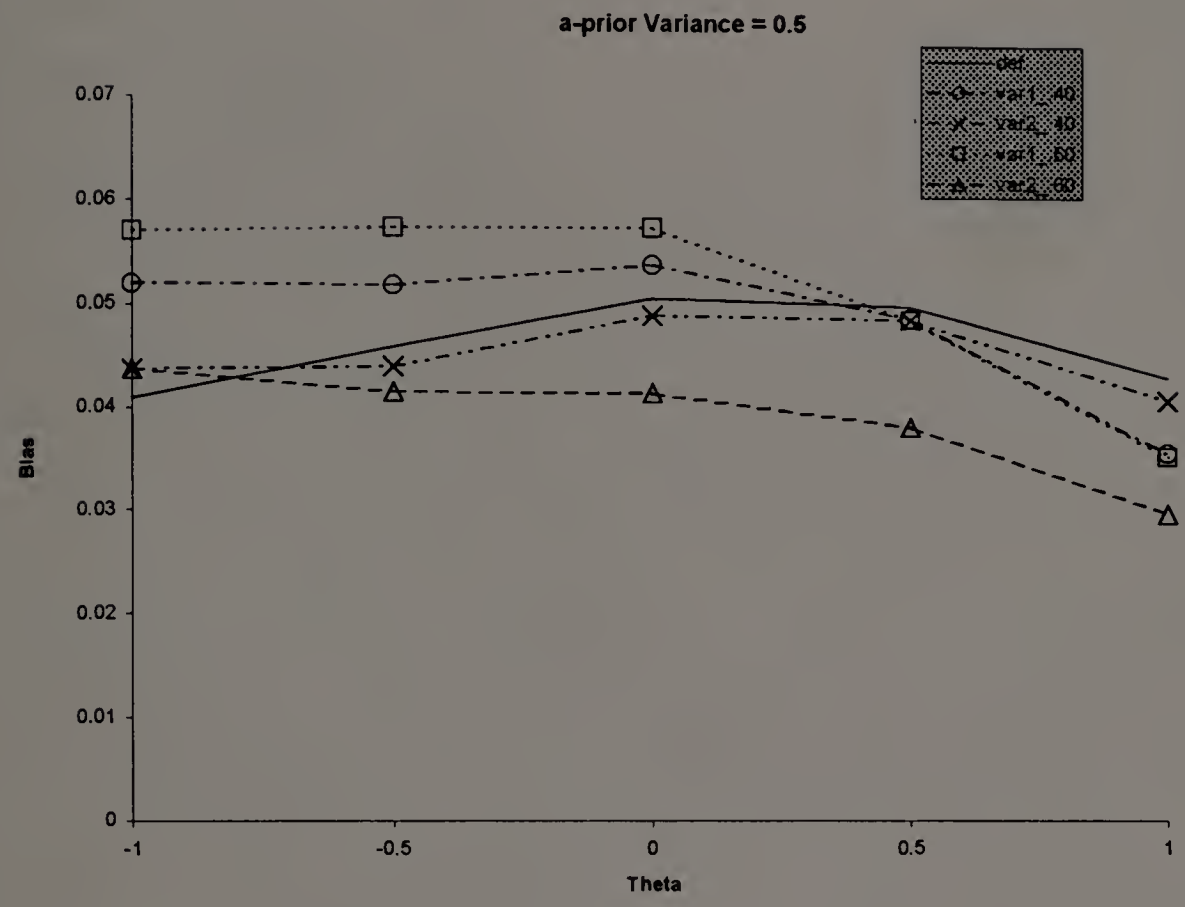


Figure 4.16

Bias of Estimated Information Functions

N=1000, n=15



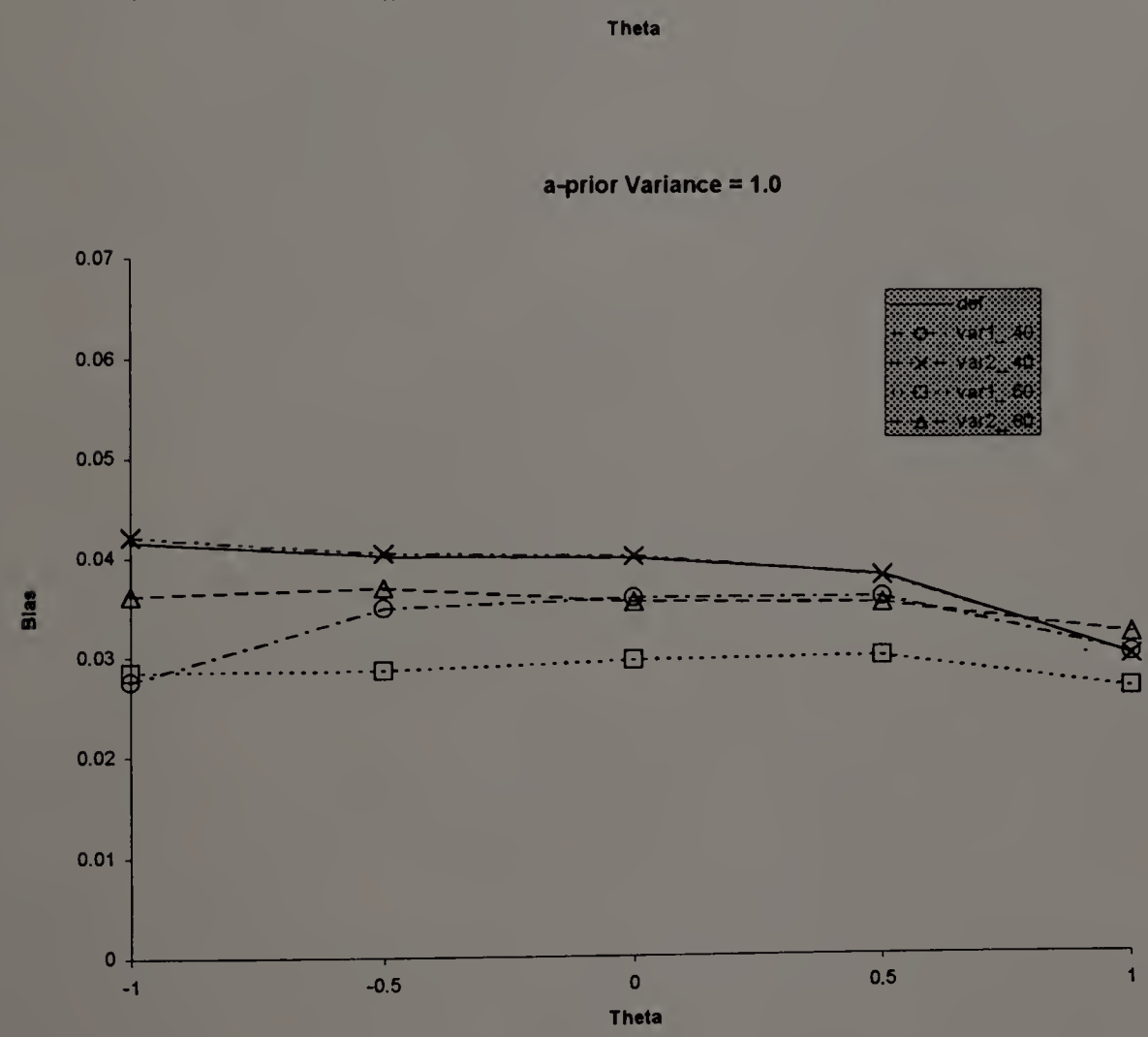
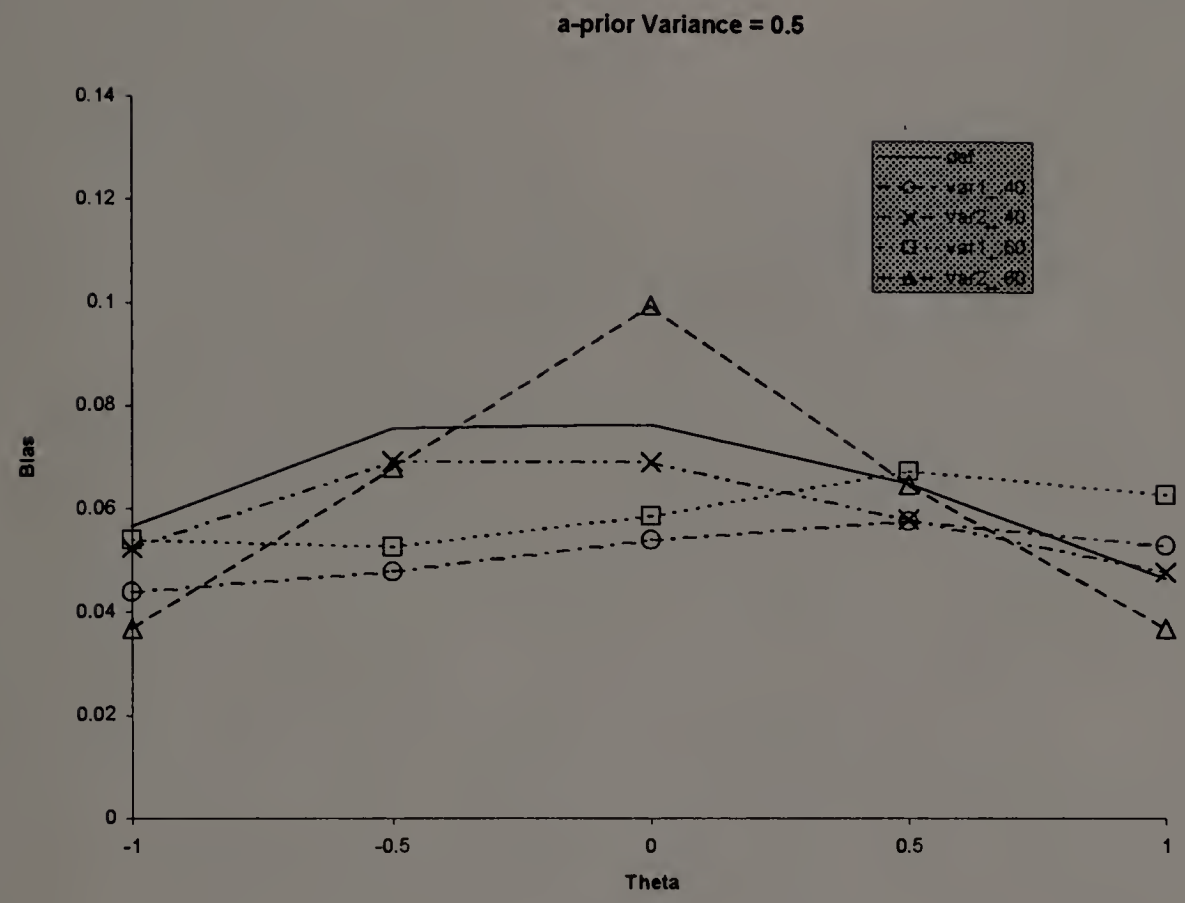


Figure 4.17

Bias of Estimated Information Functions

N=100, n=25



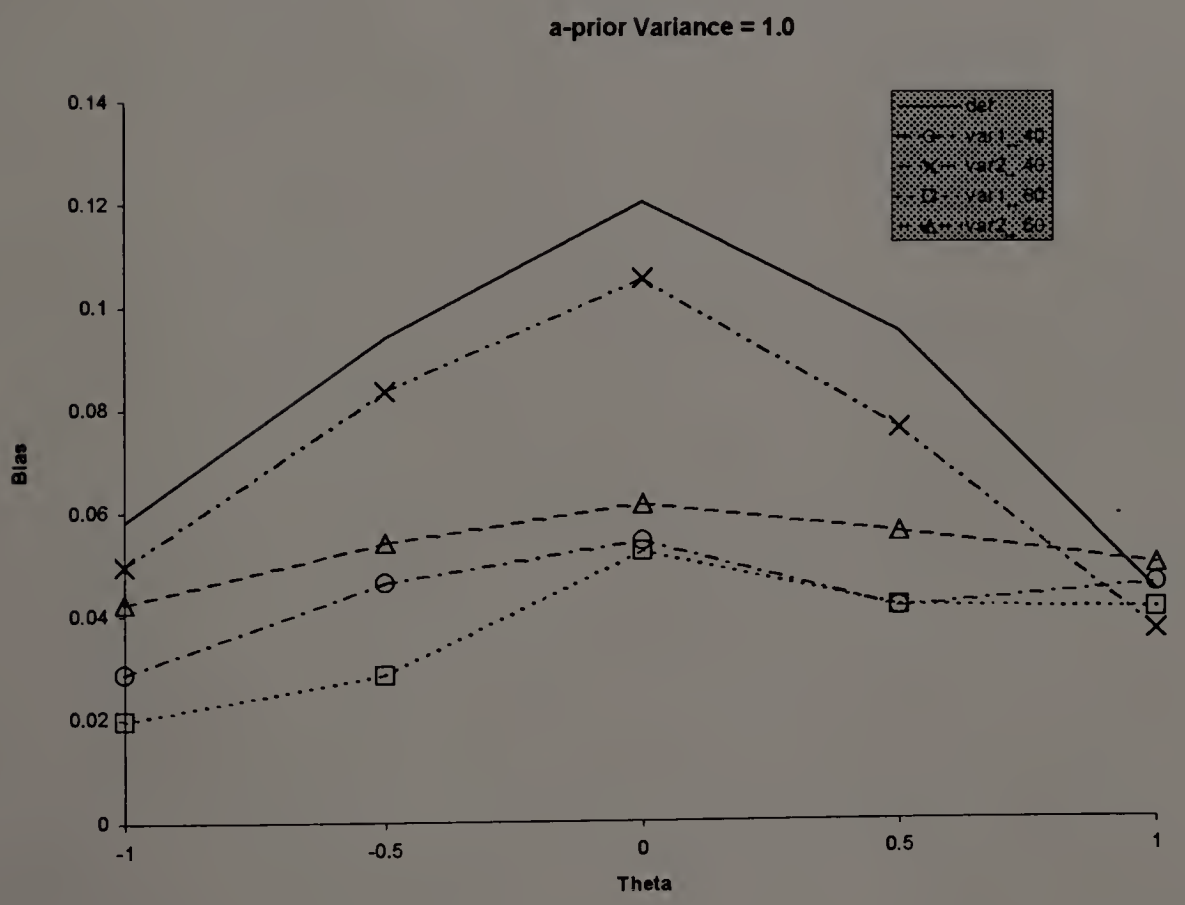
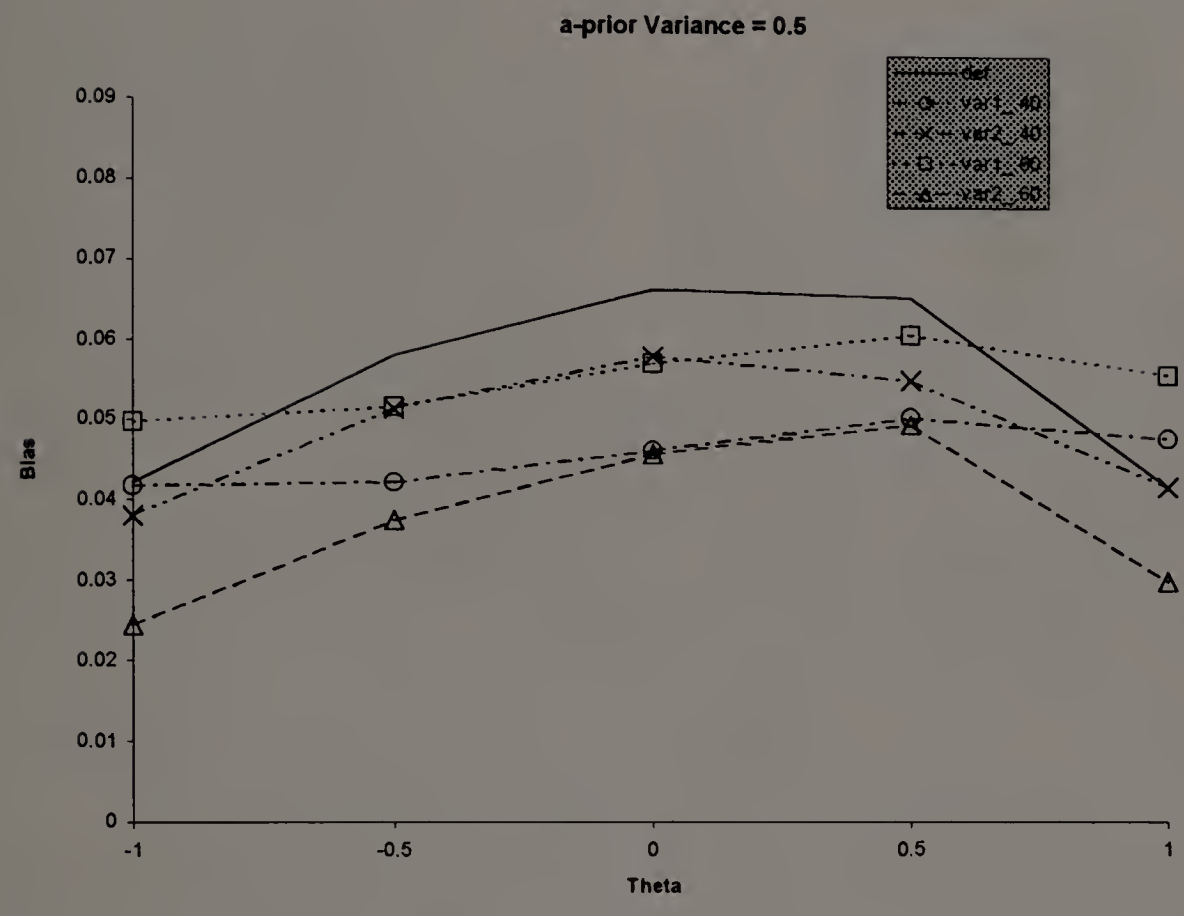


Figure 4.18

Bias of Estimated Information Functions

N=200, n=25



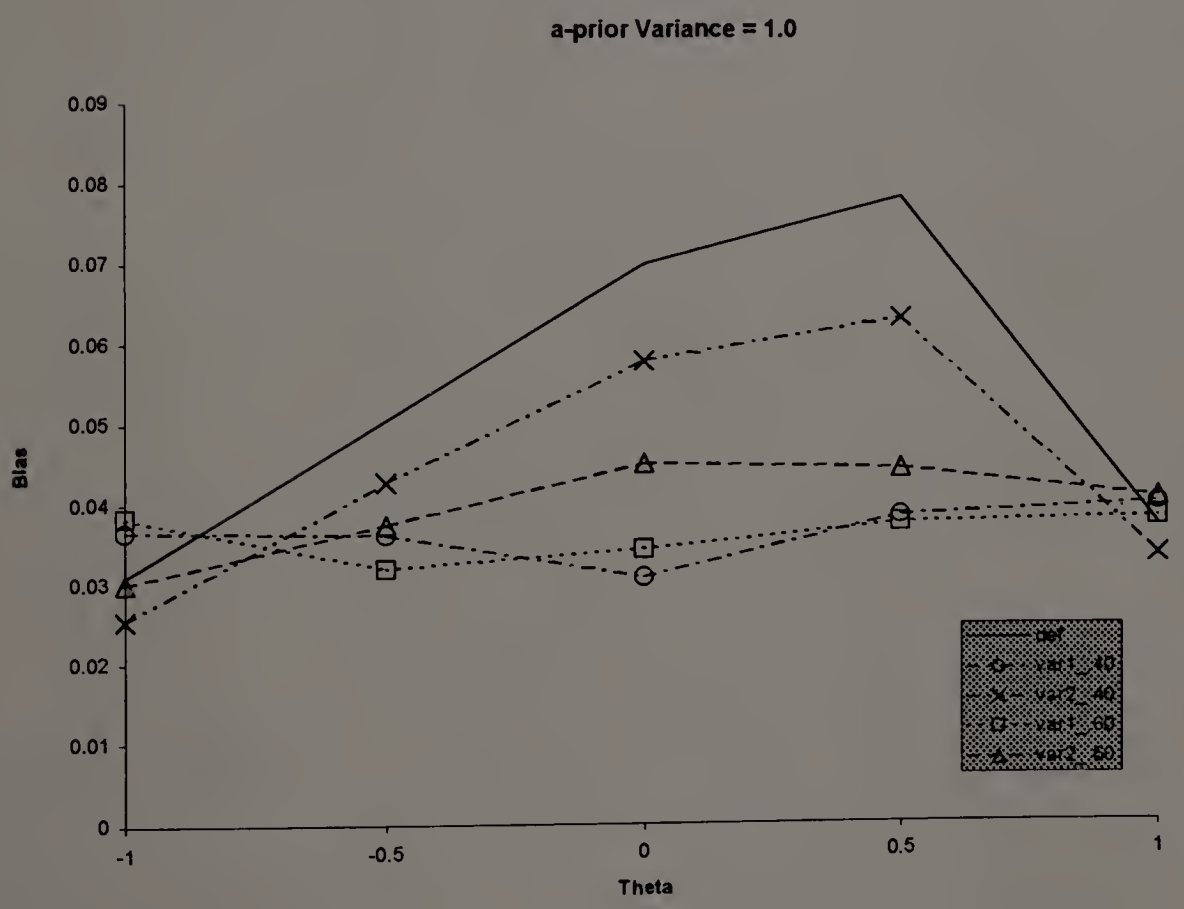
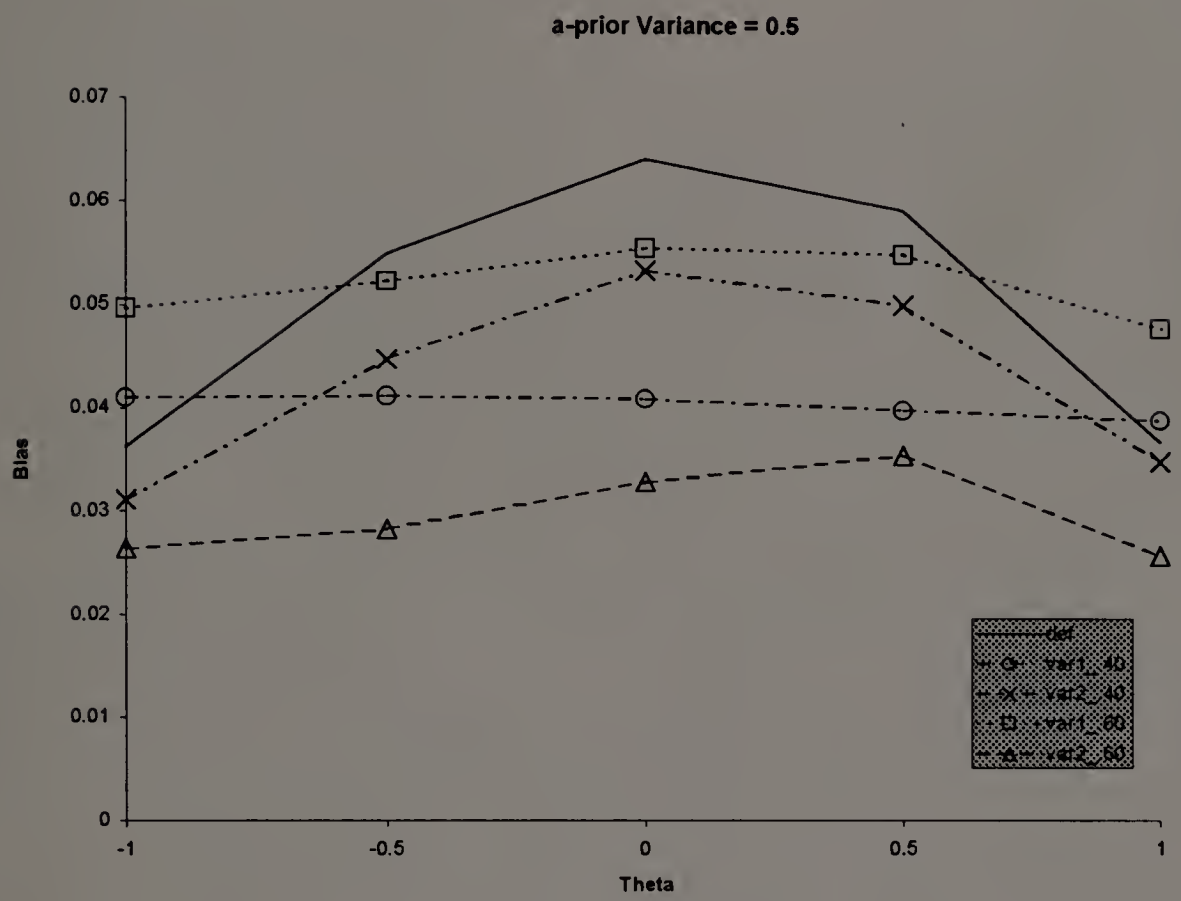


Figure 4.19

Bias of Estimated Information Functions

N=500, n=25



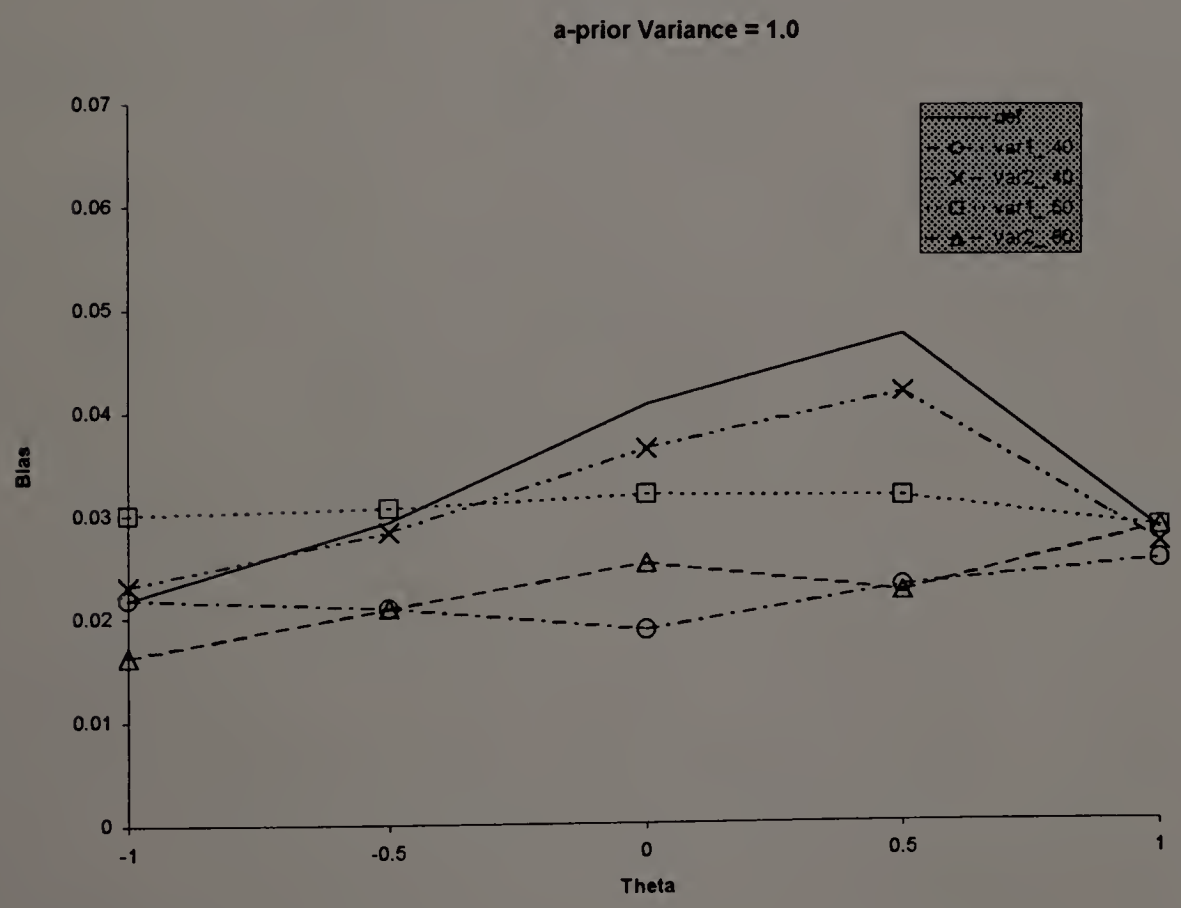
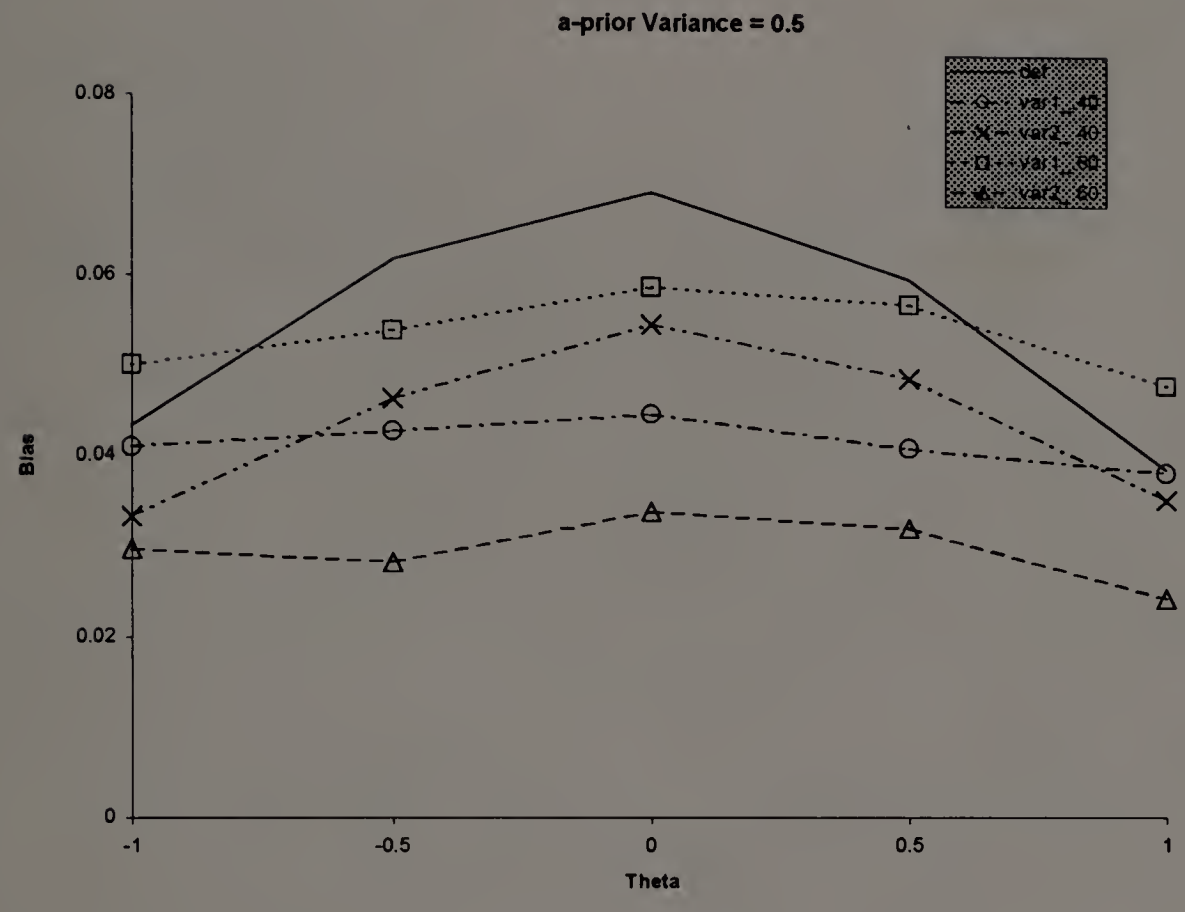


Figure 4.20

Bias of Estimated Information Functions

N=1000, n=25



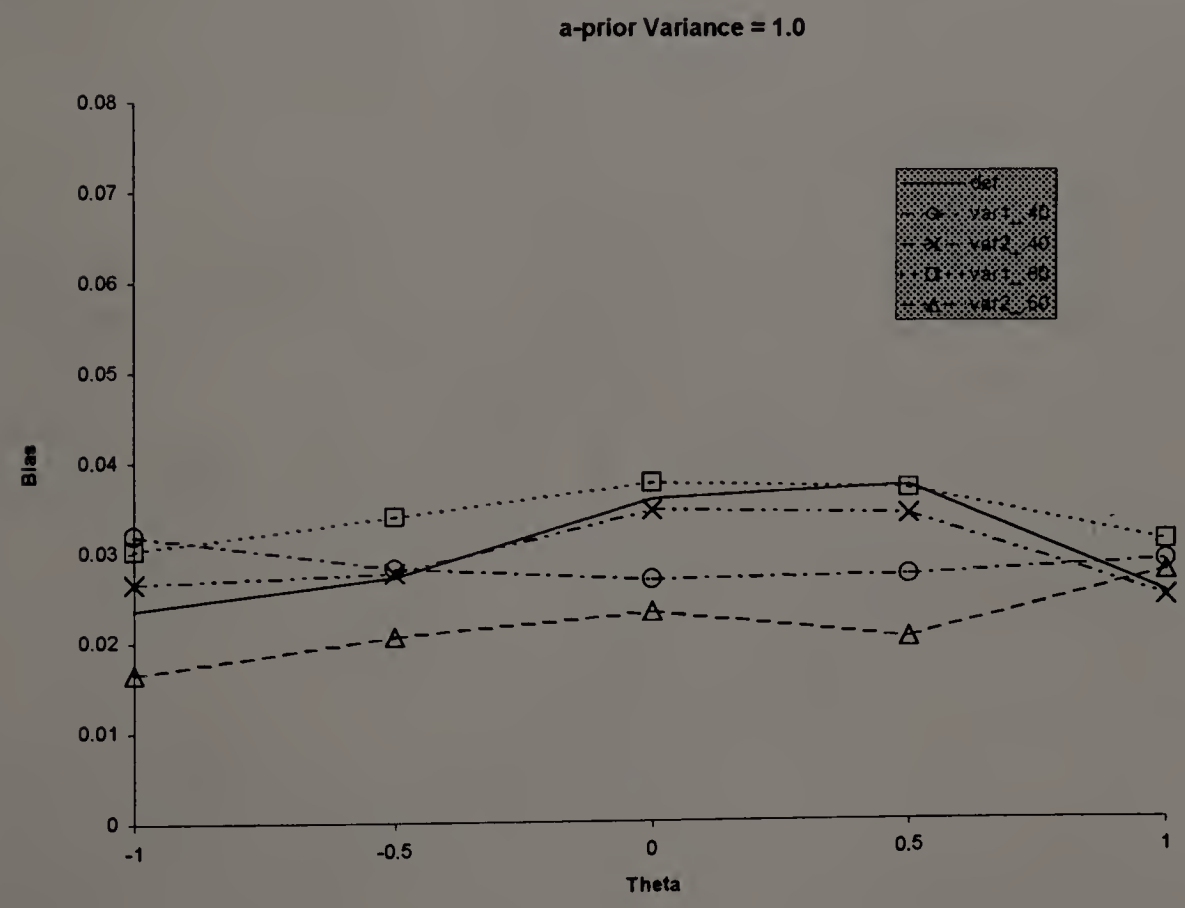
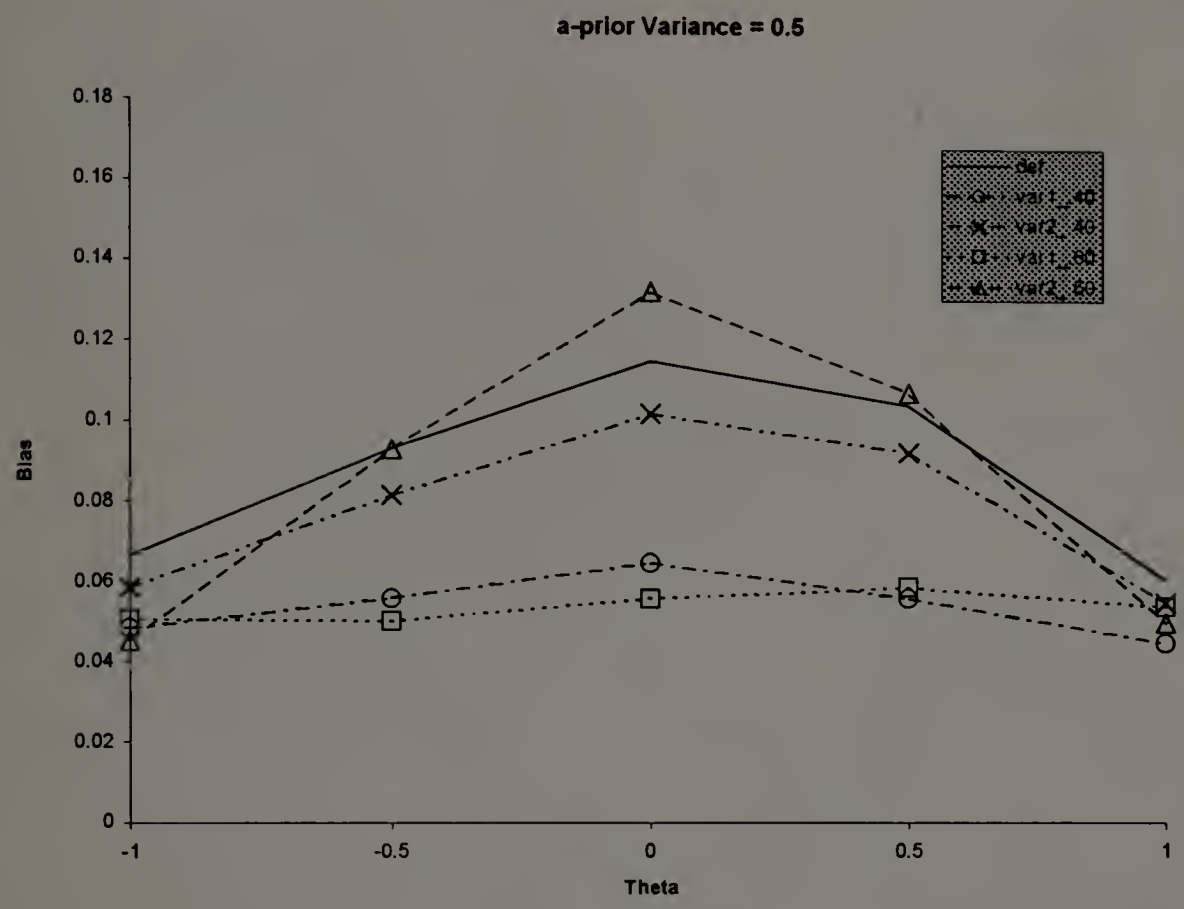


Figure 4.21

Bias of Estimated Information Functions

N=100, n=40



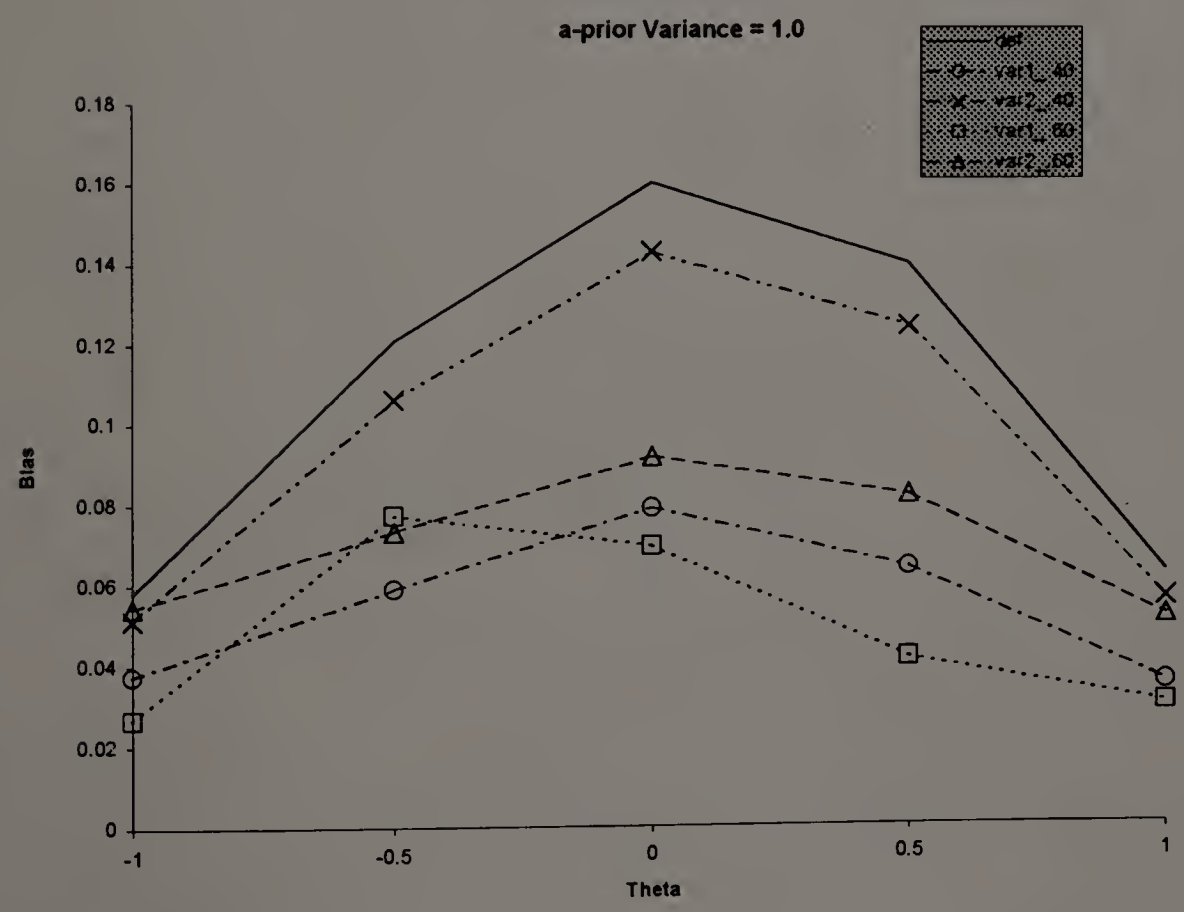
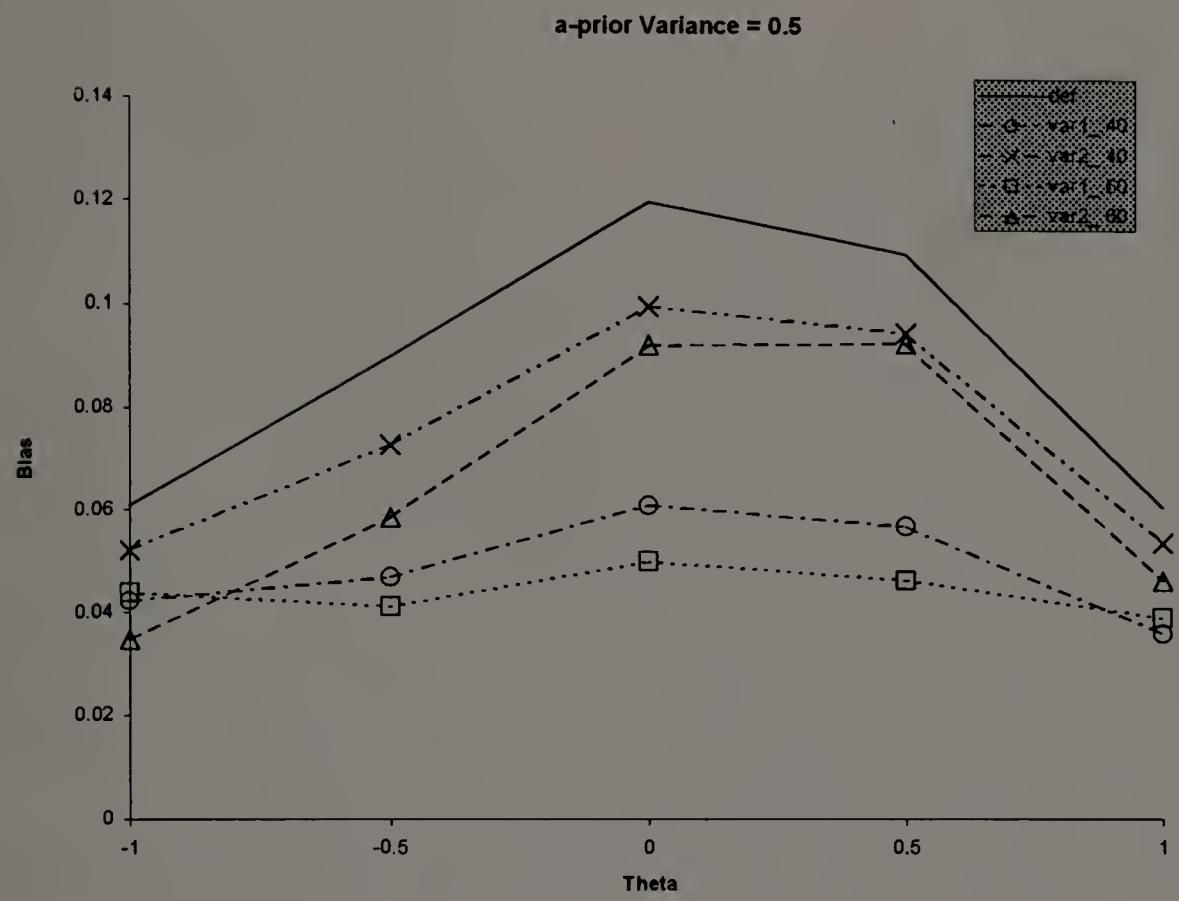


Figure 4.22

Bias of Estimated Information Functions

N=200, n=40



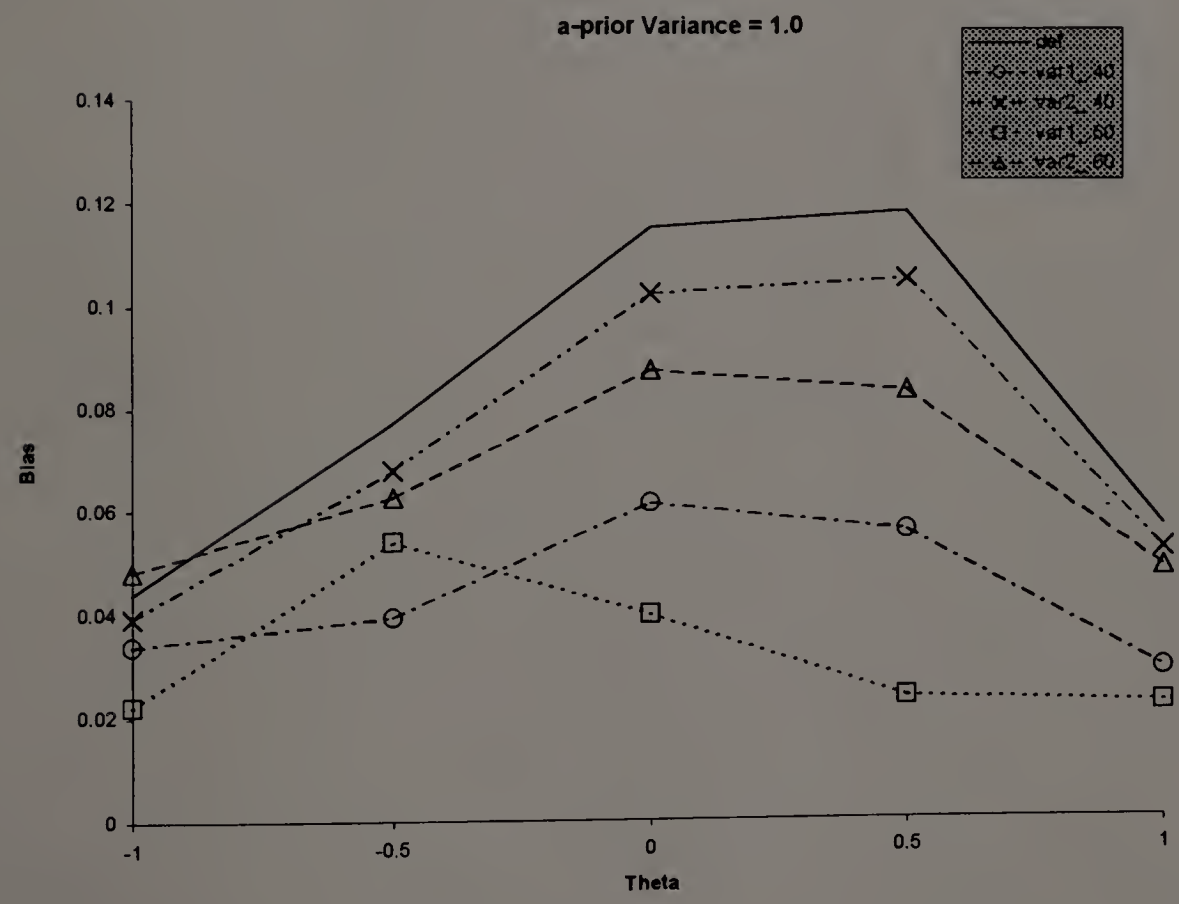
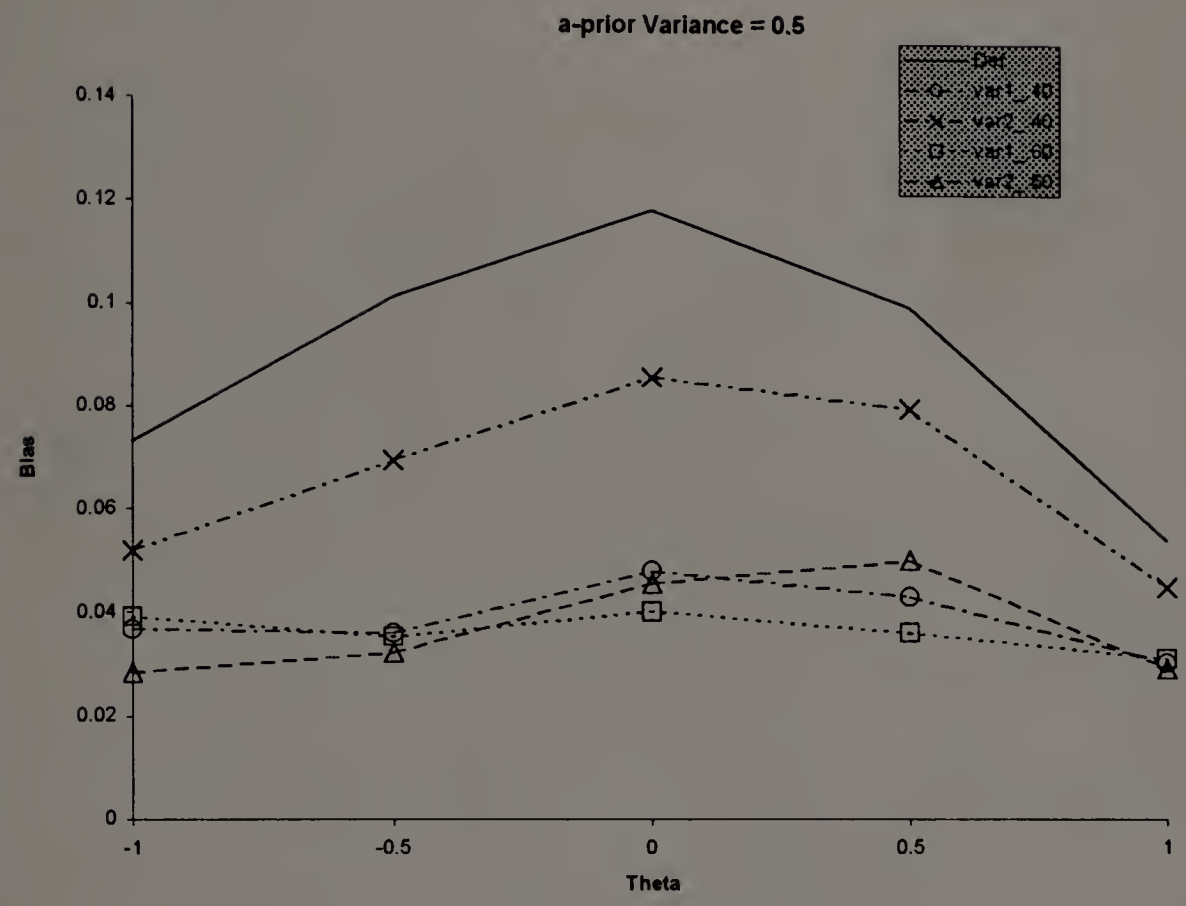


Figure 4.23

Bias of Estimated Information Functions

N=500, n=40



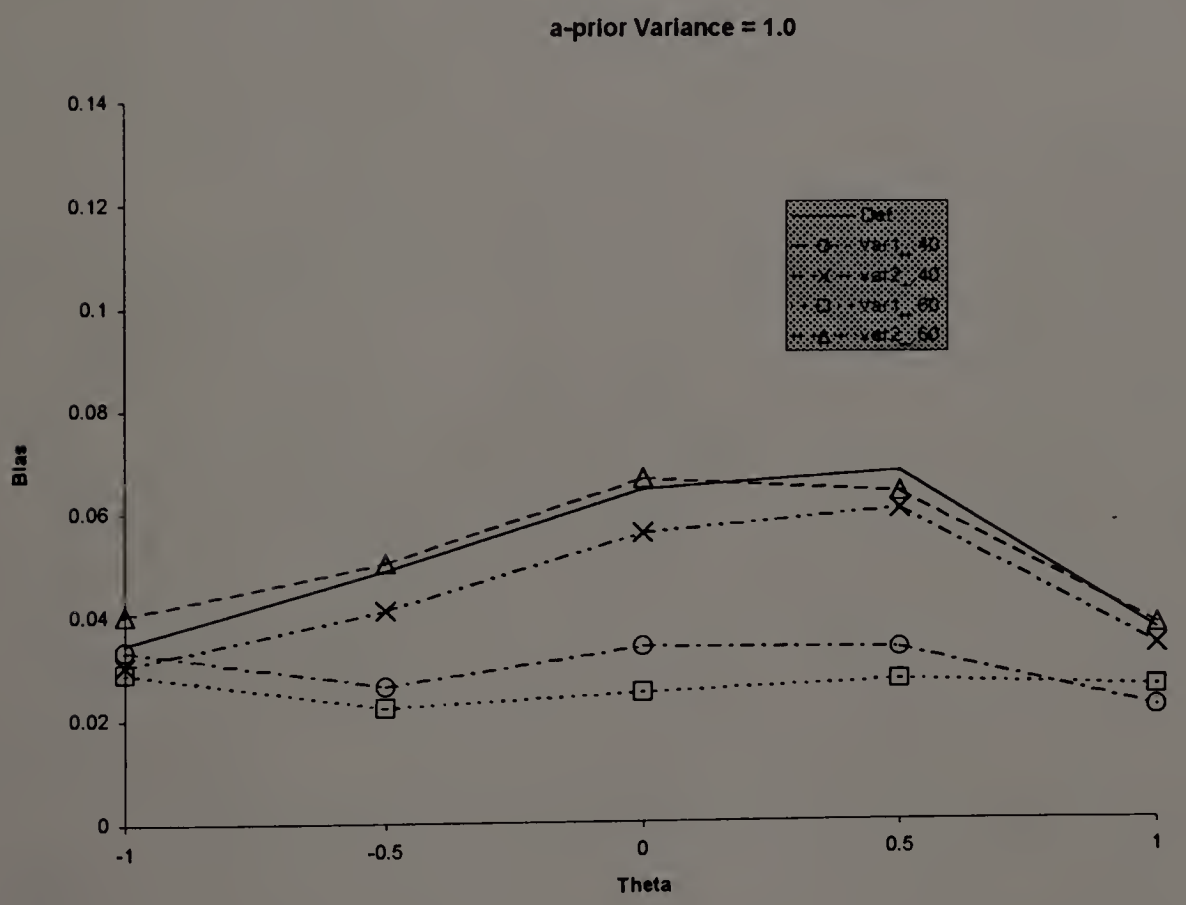
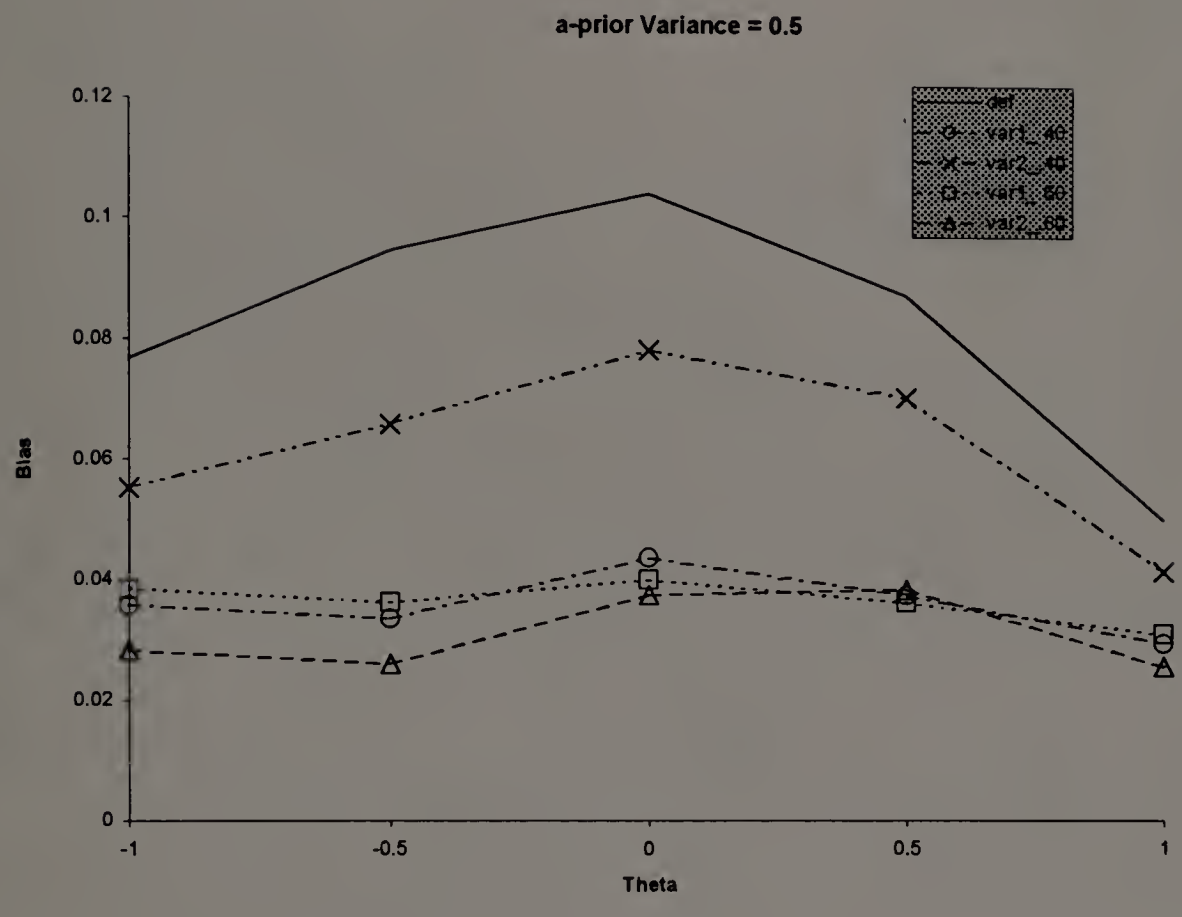
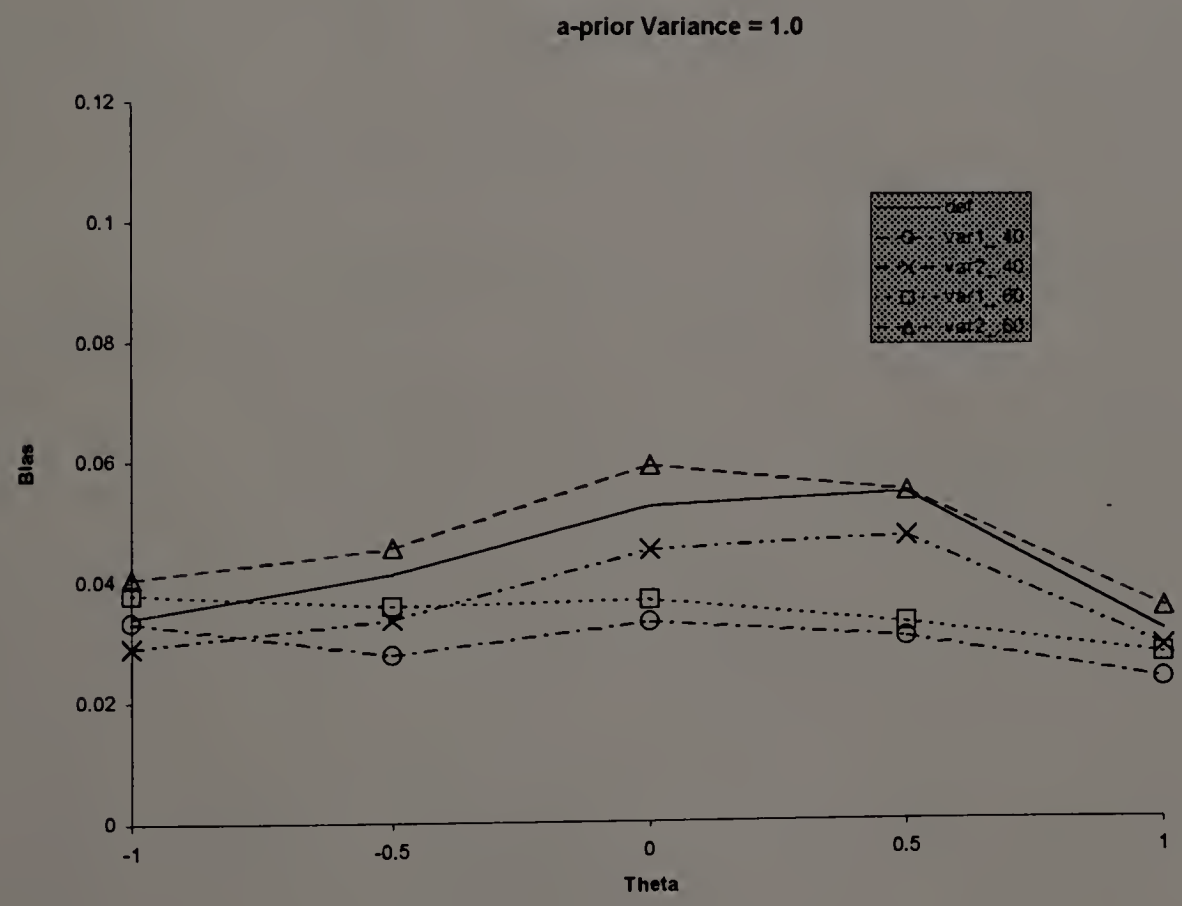


Figure 4.24

Bias of Estimated Information Functions

N=1000, n=40





CHAPTER 5

SUMMARY AND CONCLUSIONS

The previous chapter reported detailed results of the study. This chapter includes a summary of the findings, the significance of the findings, the delimitations of the study and directions for future research.

5.1 Summary of Findings

5.1.1 Summary of Item Parameter Recovery Results

The purpose of this study was to investigate the feasibility of using collateral information about items to improve the estimation of the item parameters in the three-parameter logistic model in the situation where only a small sample is available for calibration. The idea was to improve item estimation by including information that could be easily obtained about the items in the estimation process. By setting item-specific priors based on this collateral information, the more informed priors would enable the accurate estimation of item parameters in the absence of large numbers of examinees. By reducing the requisite sample sizes, the calibration of pretest items in a CAT environment would be enhanced. Different levels of collateral information were specified (simulated by changing the correlation between the parameter and the predictor) to try to determine how much information would be necessary to see improvements in both the estimation of the item parameters, as well as the estimation of the item information functions.

Regarding the estimation of the item parameters, incorporating the collateral information into the estimation lead to improvements in estimation for both the *a*- and the *b*-parameters, regardless of the criterion considered. No improvement in estimation was obtained for the *c*-parameter. However, there was no one prior that leads to the greatest improvements in estimating both the *a*- and *b*-parameters. In fact, for a given parameter, no prior lead to the most improvement based on all criteria: RMSE, bias, SD. This result is not surprising. For example, in estimating the *a*-parameter, the more informative prior on the *a*-parameter lead to more accurate estimation (in terms of RMSE), however a less informative prior lead to less biased estimates. It would be expected that the more informative prior would produce more biased estimates. Furthermore, it would be expected that more informative prior would lead to the greatest accuracy in estimation (provided it is appropriate), thus this results is not surprising. However, it is important to note that while no one prior could be superior in all cases, all item-specific priors did lead to improvement over the default priors in BILOG. Therefore, although RMSE, bias and SD cannot be minimized for all parameters simultaneously, choosing to minimize one does still result in improvements in the others. Additionally, the greatest improvements in estimation did occur for the smaller samples. Again, this is not a surprising result.

While there is no prior that is superior to the others in all cases, it is still possible to make some recommendations. In deciding on a prior distribution for the *a*-parameter, if the prior that minimizes the RMSE is chosen, the bias is still greatly reduced, while if the prior that minimizes the bias is selected, the RMSEs do not necessarily remain as small. Therefore, for the maximum benefits on all criteria, the

priors that reduce the RMSE the most are preferred. More specifically, the default prior should be used on the *a*-parameter, and, generally speaking, the most informative prior for the *b*-parameter ($\rho = .60$ and the smaller variance), except when a small number of items is used, in which case the more informative prior resulting from the lower correlation should be used. For the *b*-parameter, the recommendations are different.

When the *b*-parameter is of interest, for the larger sample sizes (N=500, 1000) the bias that produces the smallest RMSE also produces the least biased estimates. The prior for the *a*-parameter is the less informative prior, and the prior for the *b*-parameter is the most informative prior: the informative prior based on $\rho = .60$. In the small sample sizes, however, this is not the case. The prior that produces the best results in terms of RMSE and bias is the prior that minimizes the bias. In this case, the prior that minimizes the RMSE leads to estimates that are much more biased. The guidelines in this case are much more straightforward. In the smaller sample cases (N= 100, 200), the informative prior should be placed on the *a*-parameter, and the less informative prior resulting from the correlation of .60 should be used.

Unfortunately, there is no prior that maximizes the estimation of both the *a*- and *b*-parameters. However, as mentioned, the estimation of individual parameters is often not the main interest. For example, in this study, the aim is to improve item parameter estimation so as to improve item selection in CAT, therefore, in this study the examination of the item information function was examined.

5.1.2 Summary of Item Information Recovery Results

Perhaps more important than the recovery of the item parameters is the recovery of the item information function. Estimating the item information function uses all three parameters simultaneously. Therefore, the improvements for individual parameters can be combined. In recovering the item information function, both the RMSE and bias of the estimates were considered. In the case of the item parameters, there was no single prior that minimized the RMSE, bias and SD across parameters. In the case of estimating item information functions, there is about 50% overlap between the prior that minimizes the RMSE as well as the bias. In most cases, the differences are in the specification of the prior variance for the *b*-parameter. In the case where there is not agreement on which prior is best, the prior that yields the smallest bias still leads to sizable improvements in RMSE. However, the priors that yield the smallest RMSE do not always yield estimates that contain acceptable levels of bias. Therefore, to jointly minimize RMSE and bias, the prior that minimizes the bias should be prioritized.

Although there were some inconsistencies, based on the results of this study some general recommendations can be made regarding the specification of prior distributions to maximize the recovery of the information function, both in terms of accuracy and bias. Placing a less informative prior on the *a*-parameter leads to the best results. The default prior in BILOG for the *a*-parameter may be too informative. Considering the reduction in bias of the *a*-parameter when using a less informative prior, it is not surprising that it also leads to the best recovery of the information function. In terms of the *b*-parameter, for small samples, the better the collateral

information on the *b*-parameter, the better the estimation of the information function. Further, with a small sample size a more informative *b*-prior would be recommended, but as sample size increases, so should the prior variance. Again, these conclusions are sensible, as the less data available, the more the prior distribution helps. As sample size increases, the data provide more information themselves, leading to less dependence on the prior for estimation.

The more accurate recovery of the item information function is not surprising given the results for the *a*-parameter. In calculating the information function, it is necessary to square the *a*-parameter. Therefore, as the *a*-parameters are difficult to estimate, the resulting error becomes magnified. By reducing the error in the *a*-parameter, even modestly, the effects on the recovery of the information function are greater.

The results presented here suggest that the use of collateral information about item parameters does lead to improvement of estimating item parameters and item information functions, especially in small samples. One contradictory finding is in the estimation of the *c*-parameter. The work of Swaminathan et al. (in press) showed that the used of item-specific priors lead to a decrease in the error of estimation of all parameters, most notably in the *a*- and *c*-parameters. Further work in this area is necessary in order to resolve the differences in the two studies.

5.2 Significance of Results

The results of this study suggest that the current practice of estimating item parameters without the use of available collateral information should be

reconsidered. By incorporating additional information about the item parameters in the estimation process, improvements in estimation of the item parameters, and hence the item information functions, are found for all sample sizes, and primarily small samples. Most notably, incorporating the auxiliary information reduces the bias of the estimates of the *a*-parameters, leading to a reduction of bias in the information function. As mentioned previously, this is especially important in the CAT environment, where information plays the primary role for item selection, via the *a*-parameter. Since the amount of information in an item is proportional to the square of the *a*-parameter, choosing the item with the highest *a*-parameters is equivalent to choosing the most informative items. By decreasing the bias in the estimates, the items will not be selected merely because of the poor estimates of the *a*-parameter, but because the items are more highly discriminating, leading to more accurate estimates of ability, and more accurate estimates of the standard error of the resulting estimates.

While not a technological advancement, the ease of implementation of the procedure outlined here deserves mention. The technique described here is easy to implement with existing software, and relies only on existing information. Especially in a CAT setting, the amount of information available on specific items may be substantial. Response time has been shown to be a good predictor of item difficulty, and in a CAT this information is routinely collected. Therefore, what is proposed here is practical for any testing organization that has an operational CAT system. This feature should not be under appreciated, as many advances never get implemented, as the effort required to do so is either cost or time prohibitive. So while the results

might not warrant a lot of additional work, improvements of this magnitude are certainly worth the small effort required to obtain them.

5.3 Delimitations and Directions for Future Research

The findings of this study are limited by several factors. First, and foremost, the study is based on generated collateral information as well as items, examinees, and the associated response data. The value of a simulation study is that truth is known, which allows one to determine how well a given procedure performs. However to the extent that the generated data does not mirror reality, the results are limited. Every effort was made to generate data that was as realistic as possible, but the generalizability is still limited. One important aspect of the generated data that will affect the generalizability of the findings is the nature of the data studied. That the examinees did not respond to the items adaptively is not so much an issue, however the use of a complete data matrix may be. In many cases, the data available for calibration is not complete, but is in the form of a sparse data matrix with a lot of missing data. Therefore, an obvious extension of the work started here is to replicate the method with both a sparse data matrix, as well as with actual data. Although the true parameter values are not known with real data, if a large enough data set is available the large-sample estimates can be used as true values.

The importance of this study is based on the fact that the reduction in error of estimates of the *a*-parameter and more specifically, the item information function, will lead to improved estimates of ability. Therefore, examining the effect of the improvements on the ability estimation is a logical next study.

The determination of the prior distribution from the collateral information is another useful avenue of research. The work of Swaminathan et al. used the information directly, and this study used linear regression for predicting item parameters. Additional methods for translating collateral information into prior distributions may lead to better results than the previous work. As an example, using Bayesian networks for predicting item parameters may lead to better predictions than the linear regression, which may in turn lead to better prior distributions, based on the same information.

While this study considered the effect of collateral information about items on item parameter estimation, a similar approach can be taken where collateral information about examinees is used as well. The information about examinees can be used to set prior for θ. Hence, the collateral information about items can improve estimates of items, which may lead to improved ability estimation, and the auxiliary information about examinees could lead directly to improved ability estimation. The combined effects of the two types of information may really improve the estimation of ability.

In addition to alternatives for determining the prior distributions, the method for estimation may also lead to different results. Estimates in this case where obtained using BILOG, and are MAP estimates. Using other estimation techniques, such as Markov Chain Monte Carlo methods may provide better estimates. Among the differences in the procedures is that the estimates obtained are EAP estimates, which theoretically minimize the mean square error. Hence, using the MCMC estimates may produce even better results when the collateral information is introduced. Although

the success of estimating the item parameters using MCMC techniques have not shown to improve estimation in the traditional cases, when sparse data matrices are considered, more success has been shown.

Given the success in improving the item parameter estimates using small samples for the three parameters model, extending the methods proposed here to the polytomous models would be of great interest. In the polytomous case, even in large samples, there are often response categories that are infrequently used, causing for poor calibration of the threshold parameters for those categories. A common response is to collapse response categories, as adequate estimates cannot be retained. While this solves the problem of estimation, it reduces the amount of information available, hence decreasing the accuracy of ability estimation. If the types of methods proposed here can be used to aid in the estimation of those category parameters, then the information for each category need not be eliminated.

The number of studies that can be undertaken from this point is limitless. As in any study, the specific factors that were manipulated as well as the levels of those factors was a decision that could have been made differently. Had different sample sizes, test lengths, correlations, or prior variances been chosen, the results may have been different. Further, other factors may have been chosen. For instance, changing the mean of the prior on the *a*-parameter would likely have an effect on estimation as would placing item-specific priors on the *a*-parameters as well. The study represents a first foray into the realm of using predicted item parameters to set item-specific priors on parameters. As such, there are many limitations. However, pursuing these

additional lines of study can help inform the area of item parameter calibration and pretesting in a CAT environment.

APPENDIX

ADDITIONAL TABLES

Table A.1

Average RMSE of *c*-parameter

a Prior	ρ	*b* Prior	Sample Size 100	200	500	1000
Test Length = 15						
Var = 1	.40	Default	0.194	0.201	0.214	0.214
		Var 1	0.190	0.195	0.202	0.202
		Var 2	0.191	0.198	0.210	0.208
	.60	Var 1	0.194	0.195	0.204	0.199
		Var 2	0.190	0.196	0.205	0.203
Var = .5	.40	Default	0.175	0.180	0.190	0.187
		Var 1	0.189	0.191	0.197	0.194
		Var 2	0.179	0.182	0.191	0.188
	.60	Var 1	0.194	0.186	0.205	0.203
		Var 2	0.182	0.184	0.193	0.189
Test Length = 25						
Var = 1	.40	Default	0.195	0.205	0.211	0.217
		Var 1	0.190	0.200	0.201	0.201
		Var 2	0.192	0.200	0.205	0.209
	.60	Var 1	0.206	0.210	0.201	0.197
		Var 2	0.187	0.198	0.202	0.204
Var = .5	.40	Default	0.178	0.178	0.177	0.178
		Var 1	0.191	0.198	0.192	0.189
		Var 2	0.180	0.182	0.180	0.180
	.60	Var 1	0.196	0.208	0.206	0.205
		Var 2	0.185	0.187	0.183	0.182
Test Length =40						
Var = 1	.40	Default	0.198	0.212	0.220	0.217
		Var 1	0.189	0.201	0.202	0.201
		Var 2	0.194	0.206	0.213	0.211
	.60	Var 1	0.207	0.212	0.199	0.194
		Var 2	0.189	0.201	0.205	0.205
Var = .5	.40	Default	0.177	0.186	0.184	0.182
		Var 1	0.195	0.202	0.193	0.189
		Var 2	0.181	0.189	0.185	0.183
	.60	Var 1	0.196	0.207	0.204	0.202
		Var 2	0.189	0.192	0.188	0.184

Table A.2

Average Absolute Bias of *c*-parameter

a Prior	ρ	*b* Prior	Sample Size 100	200	500	1000
Test Length = 15						
Var = 1	.40	Default	0.099	0.101	0.095	0.088
		Var 1	0.102	0.104	0.099	0.092
		Var 2	0.100	0.101	0.096	0.088
	.60	Var 1	0.103	0.108	0.104	0.096
		Var 2	0.108	0.112	0.109	0.102
Var = .5	.40	Default	0.108	0.113	0.111	0.105
		Var 1	0.107	0.112	0.108	0.102
		Var 2	0.108	0.113	0.110	0.104
	.60	Var 1	0.103	0.100	0.111	0.104
		Var 2	0.101	0.102	0.097	0.089
Test Length = 25						
Var = 1	.40	Default	0.102	0.102	0.094	0.087
		Var 1	0.104	0.109	0.100	0.094
		Var 2	0.102	0.103	0.095	0.088
	.60	Var 1	0.106	0.116	0.108	0.101
		Var 2	0.105	0.117	0.112	0.106
Var = .5	.40	Default	0.112	0.118	0.115	0.111
		Var 1	0.110	0.117	0.112	0.106
		Var 2	0.111	0.118	0.113	0.108
	.60	Var 1	0.108	0.123	0.117	0.111
		Var 2	0.100	0.104	0.096	0.089
Test Length =40						
Var = 1	.40	Default	0.098	0.098	0.088	0.082
		Var 1	0.101	0.100	0.091	0.087
		Var 2	0.098	0.098	0.088	0.082
	.60	Var 1	0.107	0.105	0.097	0.092
		Var 2	0.105	0.111	0.106	0.101
Var = .5	.40	Default	0.110	0.117	0.113	0.107
		Var 1	0.105	0.109	0.103	0.099
		Var 2	0.108	0.114	0.109	0.104
	.60	Var 1	0.109	0.113	0.106	0.102
		Var 2	0.102	0.098	0.089	0.083

Table A.3

Average Standard Deviation of *c*-parameter

a Prior	ρ	*b* Prior	Sample Size 100	200	500	1000
Test Length = 15						
Var = 1	.40	Default	0.020	0.021	0.026	0.026
		Var 1	0.024	0.024	0.027	0.027
		Var 2	0.021	0.022	0.026	0.026
	.60	Var 1	0.025	0.025	0.029	0.029
		Var 2	0.023	0.024	0.029	0.029
Var = .5	.40	Default	0.025	0.026	0.032	0.033
		Var 1	0.023	0.024	0.027	0.028
		Var 2	0.024	0.025	0.030	0.030
	.60	Var 1	0.025	0.023	0.028	0.029
		Var 2	0.022	0.022	0.026	0.026
Test Length = 25						
Var = 1	.40	Default	0.021	0.021	0.022	0.024
		Var 1	0.025	0.027	0.026	0.026
		Var 2	0.021	0.022	0.023	0.024
	.60	Var 1	0.030	0.030	0.028	0.027
		Var 2	0.023	0.025	0.027	0.029
Var = .5	.40	Default	0.025	0.027	0.030	0.034
		Var 1	0.023	0.025	0.026	0.028
		Var 2	0.023	0.025	0.028	0.031
	.60	Var 1	0.026	0.028	0.028	0.028
		Var 2	0.023	0.023	0.023	0.024
Test Length =40						
Var = 1	.40	Default	0.021	0.024	0.025	0.026
		Var 1	0.027	0.029	0.027	0.027
		Var 2	0.022	0.025	0.026	0.026
	.60	Var 1	0.030	0.033	0.029	0.027
		Var 2	0.024	0.027	0.029	0.030
Var = .5	.40	Default	0.026	0.031	0.035	0.035
		Var 1	0.023	0.027	0.028	0.029
		Var 2	0.025	0.028	0.032	0.032
	.60	Var 1	0.026	0.029	0.029	0.028
		Var 2	0.024	0.026	0.026	0.026

References

Barnes, L. B., & Wise, S. L. (1991). The utility of a modified one-parameter IRT model with small samples. *Applied Measurement in Education, 4,* 143-157.

Bennett, R. E. (1999). Using new technology to improve assessment. *Educational Measurement: Issues and Practice, 18*(3), 5–12.

Dennis, I., Handley, S., Bradon, P., Evans, J., & Newstead, S. (in press). Approaches to modeling item-generative tests. In S. H. Irvine & P. Kyllonen (Eds.), *Item generation for test development.* Mahwah, NJ: Erlbaum.

Embretson, S. E. (in press). Generating abstract reasoning items with cognitive theory. In S. H. Irvine & P. Kyllonen (Eds.), *Item generation for test development.* Mahwah, NJ: Erlbaum.

Enright, M. K., & Sheehan, K. M. (in press). Modeling the difficulty of quantitative reasoning items: Implications for item generation. In S. H. Irvine & P. Kyllonen (Eds.), *Item generation for test development.* Mahwah, NJ: Erlbaum.

Enright, M. K., Morley, M., & Sheehan, K. M. (1999). *Items by design: The impact of systematic feature variation on item statistical characteristics* (Educational Testing Service Research Report no. RM-99-20). Princeton, NJ: Educational Testing Service.

Gifford, J. A., & Swaminathan, H. (1990). Bias and the effect of priors in Bayesian estimation of parameters of item response models. *Applied Psychological Measurement, 14*(1), 33-43.

Hambleton, R. K., & Jones, R. W. (1994). Item parameter estimation errors and their influence on test information functions. *Applied Measurement in Education, 7,* 171-186.

Hambleton, R. K. &. & Swaminathan, H. (1985). *Item response theory principles and applications.* Boston, MA: Kluwer-Nijhoff Publishing.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory.* Newbury Park, CA: Sage Publishing.

Harwell, M.R., & Baker, F. B. (1991). The use of prior distributions in marginalized Bayesian item parameter estimation: A didactic. *Applied Psychological Measurement, 14,* 4, 375-389.

Harwell, M. R., & Janosky, J. E. (1991). An empirical study of the effects of small datasets and varying prior variances on item parameter estimation in BILOG. *Applied Psychological Measurement, 15,* 279-291.

Hornke, L. F. (in press). Item generation models for higher order cognitive functions. In S. H. Irvine & P. Kyllonen (Eds.), *Item generation for test development.* Mahwah, NJ: Erlbaum.

Lewis, C. (2001). Expected response functions. In A. Boomsma, M. A. J. van Duijn, &T. A. B. Snijders (Eds.), *Essays on Item Response Theory.* New York: Springer-Verlag.

Lord, F. M. (1986). Maximum Likelihood and Bayesian parameter estimation in Item Repsonse Theory. *Journal of Educational Measurement, 23,* 157-162.

Mislevy, R. J. (1986). *Exploiting auxiliary information about examinees in the estimation of item parameters.* (Research Report 86-18-ONR). Princeton, NJ: Educational Testing Service.

Mislevy, R. J. (1988). *Exploiting collateral information in the estimation of item parameters.* (Research Report NR 150-539). Princeton, NJ: Educational Testing Service.

Mislevy, R. L., & Sheehan, K. M. (1989). The role of collateral information about examinees in item parameter estimation. *Psychometrika, 54*(4), 661-679.

Mislevy R. L. & Stocking, M. L. (1989). A consumer's guide to LOGIST and BILOG. *Applied Psychological Measurement, 13,* 57-75.

Mislevy, R. J., Wingersky, M. S., & Sheehan, K. M. (1994). Dealing with uncertainty about item parameters: Expected response functions (Research Report, 94-28-ONR). Princeton, NJ: Educational Testing Service.

Neyman, J., & Scott, E.L. (1948). Consistent estimates based on partially consistent observations. *Econometrica, 16,* 1-32.

O'Hagan, A. (1976). On posterior joint and marginal modes. *Biometrika, 63,* 329-333.

Parshall, C. G., Kromrey, J. D., & Chason, W. M. (1996, June). *Comparison of alternative models for item parameter estimation with small samples.* Paper presented at the annual meeting of the Psychometric Society, Banff.

Patsula, L. N., & Pashley, P. J. (1996, April). *Pretest item analyses using polynomial logistic regression: An approach to small sample calibration problems associated with computerized adaptive testing.* Paper presented at the annual meeting of the National Council on Measurement in Education, New York.

Seong, T-J. (1990). Sensitivity of marginal maximum likelihood estimation of item and ability parameters to the characteristics of the prior ability distributions. *Applied Psychological Measurement, 14*(3), 299-311.

Singley, M. K., & Bennett, R. E. (in press). Item generation and beyond: Applications of schema theory to mathematics assessment. In S. H. Irvine & P. Kyllonen (Eds.), *Item generation for test development.* Mahwah, NJ: Erlbaum.

Sireci, S. G. (1992, August). *The utility of IRT in small-sample testing applications.* Paper presented at the annual meeting of the American Psychologocal Association, Washington, D.C.

Stone, C. A., & Lane, S. (1991). Use of restricted item response theory models for examining the stability of item parameter estimates over time. *Applied Measurement in Education, 4,* 125-141.

Swaminathan, H., & Gifford, J.A. (1982). Bayesian estimation in the Rasch model. *Journal of Educational Statistics, 7,* 175-192.

Swaminathan, H. & Gifford, J. A. (1983). Estimation of item parameters in the three-parameter latent trait model. In D. Weiss (Ed.), *New Horizons in testing.* (pp. 13-30). New York: Academic Press.

Swaminathan, H., & Gifford, J.A. (1985). Bayesian estimation in the two-parameter logistic model. *Psychometrika, 50,* 349-364.

Swaminathan, H., Hambleton, R. K., Sireci, S. G., Xing, D., Rizavi, S. M. (in press). Small sample estimation in dichotomous item response models: Effect of priors based on judgemental information on the accuracy of item parameter estimates. *Applied Psychological Measurement.*

Swaminathan, H., & Gifford, J.A. (1986). Bayesian estimation in the three-parameter logistic model. *Psychometrika, 51,* 589-601.

Tsutakawa, R. K., & Lin, H. Y. (1986). Bayesian estimation of item response curves. *Psychometrika, 51*, 251-267.

Zeng, L. (1997). Implementation of marginal Bayesian estimation with four-parameter beta prior distributions. *Applied Psychological Measurement, 21,* 143-156.